

DOCUMENT RESUME

ED 201 641

TM 810 086

AUTHOR Clark, John L. D.; Swinton, Spencer S.
 TITLE An Exploration of Speaking Proficiency Measures in the TOEFL Context. Report 4.
 REPORT NO ETS-RR-79-8
 PUB DATE Oct 79
 NOTE 79p.
 AVAILABLE FROM Educational Testing Service, Publications Order Service Section, Princeton, NJ 08541 (\$2.00 per copy).
 EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
 DESCRIPTORS Audiotape Recordings; *Communicative Competence (Languages); Correlation; *English (Second Language); Higher Education; Item Analysis; *Language Proficiency; *Language Tests; Scoring; *Speech Tests; Test Construction
 IDENTIFIERS *Foreign Service Institute Interview; Inter Rater Reliability; *Test of English as a Foreign Language; Test of Spoken English

ABSTRACT

A three-year study aimed at measuring the English speaking proficiency of Test of English as a Foreign Language (TOEFL) candidates on a direct basis is described. A variety of "semi-direct" testing formats (involving test stimuli delivered by tape recording and accompanying booklet, with examinee speech tape recorded for evaluation) were developed as part of the study. These were administered to undergraduate and graduate level non-native English speaking students on a small-scale clinical basis and, following revisions, to a fairly large (N=155) examinee group. A Foreign Service Institute (FSI)-type interview was also administered to all study participants as an external criterion measure of speaking proficiency. Correlational and scalar analyses were used to identify subsets of item formats and individual items having satisfactory correlations with the FSI criterion interview and also meeting relevant linguistic and administrative requirements. These were grouped into a prototype "Test of Spoken English" (TSE) that was found to increase predictability of total FSI variance by approximately 8% when added to score data provided by the regular (paper-and-pencil) TOEFL. (Author/RL)

 * Reproductions supplied by EDRS are the best that can be made *
 * from the original document. *

TOEFL

Research Reports



REPORT 4
OCTOBER 1979

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official NIE
position or policy.

AN EXPLORATION OF SPEAKING PROFICIENCY MEASURES IN THE TOEFL CONTEXT

John L. D. Clark
Spencer S. Swinton

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

H.C. Weidenmiller

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."



EDUCATIONAL TESTING SERVICE

ED201641

TM 810086

The Test of English as a Foreign Language (TOEFL) was developed in 1963 by a National Council on the Testing of English as a Foreign Language which was formed through the cooperative effort of over thirty organizations, public and private, that were concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board assumed joint responsibility for the program and in 1973 a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; Graduate Record Examinations Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a Policy Council that was established by, and is affiliated with, the sponsoring organizations. Members of the Policy Council represent the College Board and the Graduate Record Examinations Board and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

An Exploration of Speaking Proficiency Measures
in the TOEFL Context

John L. D. Clark
Spencer S. Swinton

Educational Testing Service
Princeton, N.J.

RR 79-8

Copyright © 1979 by Educational Testing Service. All rights reserved.

Unauthorized reproduction in whole or in part is prohibited.

Abstract

This report describes a three-year study involving the initial development and experimental administration of a variety of test formats and item types aimed at measuring the English speaking proficiency of TOEFL candidates or other non-native speakers of English on a direct basis, i.e., through the evaluation of actual speech production on the part of the examinee.

Although practical constraints of the TOEFL program ruled out of consideration, for operational purposes, face-to-face testing procedures of the Foreign Service Institute (FSI) type, a variety of "semi-direct" testing formats (involving test stimuli delivered by tape recording and accompanying booklet, with examinee speech tape recorded for evaluation) were developed as part of the study. These were administered to undergraduate and graduate level non-native English speaking students on a small-scale clinical basis and, following revisions, to a fairly large (N=155) examinee group. An FSI-type interview was also administered to all study participants as an external criterion measure of speaking proficiency.

Correlational and scalar analyses were used to identify subsets of item formats and individual items having satisfactory correlations with the FSI criterion interview and also meeting relevant linguistic and administrative requirements. These were grouped into a prototype "Test of Spoken English" (TSE) that was found to increase predictability of total FSI variance by approximately 8% when added to score data provided by the regular (paper-and-pencil) TOEFL. Further modifications in TSE content and follow-up administration with larger and more heterogeneous candidate groups are expected to increase this value.

Two additional related studies are currently under way, involving pilot test administration at regular TOEFL sites and use-validation of the TSE as a measure of the "classroom communication" effectiveness of non-native English speaking teaching assistants in U.S. academic settings.

TABLE OF CONTENTS

	<u>Page</u>
BACKGROUND OF THE STUDY.....	1
General Considerations in Study Design.....	2
THE FSI CRITERION INTERVIEW.....	4
Preliminary FSI Reliability Study.....	7
Intrarater Reliability.....	9
Interrater Reliability.....	9
Correlations with External Criteria.....	10
Considerations in Adoption of FSI Criterion.....	12
DEVELOPMENT OF INITIAL EXPERIMENTAL TEST FORMS.....	12
Final Content of Initial Test.....	26
DEVELOPMENT OF SCORING PROCEDURES.....	26
CLINICAL ADMINISTRATION OF INITIAL TEST FORMS.....	28
Content Outline of Revised Test Forms.....	34
LARGE-SCALE ADMINISTRATION OF REVISED TEST FORMS.....	34
Scoring Procedures.....	38
DATA ANALYSIS AND RESULTS.....	39
Development of Covariance Matrix.....	40
Item-Type Analyses.....	44
Individual Item Selection and Scaling.....	49
FURTHER RESEARCH AND DEVELOPMENT ACTIVITIES.....	53
REFERENCES.....	55
APPENDICES	
A. Initial Speaking Test Scoring Key.....	57
B. Language Background/Self-Appraisal Questionnaire.....	61
C. FSI Level Descriptions.....	63
D. FSI "Grid" Rating Form.....	65
E. FSI "Descriptive Statements" Rating Criteria.....	66
F. Revised Speaking Test Scoring Key.....	67

LIST OF TABLES

Table	Page
1. Intrarater Reliabilities for Long and Short Interviews.....	9
2. Interrater Reliabilities for Long and Short Interviews.....	10
3. Intercorrelations of Long and Short Interviews and Other Indices of Language Competence.....	11
4. Item Type Content of Initial Test Forms.....	27
5. Item Type Content of Revised Test Forms.....	35
6. Means and Standard Deviations for Criterion and Background Variables.....	42
7. Intercorrelations of TOEFL, FSI, and Experimental Speaking Test Scores.....	43
8. Median Correlations of Experimental Speaking Test Items with FSI and with TOEFL II+III.....	46
9. Correlations of Prototype TSE Items with Criterion Measures....	51
10. Regression Equation Predicting FSI Score from TSE Items and Subscores.....	52

BACKGROUND OF THE STUDY

Since the beginning of the testing program in 1960, the Test of English as a Foreign Language (TOEFL) has included sections which provide for a direct assessment of both the candidate's level of listening comprehension in English and his or her reading comprehension of printed English texts. (Other TOEFL sections have, in addition, been aimed at evaluating the candidate's recognitional knowledge of vocabulary and grammatical structures.) However, the direct testing of speaking proficiency in English--direct in the sense of requiring the candidate actually to speak in the test language--has not been a part of the operational TOEFL program for a number of reasons. These include not only the cost and administrative complexity of providing an active speaking test capability at the hundreds of TOEFL administration sites worldwide but also the paucity of relevant research experience and development, within both the TOEFL program and the language testing community in general, concerning the direct evaluation of speaking proficiency on a large-scale basis.

In the absence of a direct measure of candidate speaking proficiency, attention may be turned to other kinds of testing information which are known or strongly presumed to interrelate with measured speaking proficiency. A number of studies have shown substantial correlations between tests directly measuring active speaking ability and paper-and-pencil tests of listening comprehension, reading, and writing ability. For example, in John B. Carroll's large-scale study of the foreign language proficiency of college majors in French, German, Italian, Russian, and Spanish (Carroll *et al.*, 1967), correlations ranging from .58 to .86 were obtained between the listening comprehension, reading, and writing tests of the MLA Proficiency Tests for Teachers and Advanced Students and a face-to-face speaking interview of the Foreign Service Institute (FSI) type.

In a recent study within the TOEFL program, Pike (1979) found correlations of .43 to .81 between scores on a direct speaking test modeled generally on FSI procedures and the five multiple-choice subsections of the then-current TOEFL: Listening Comprehension, English Structure, Vocabulary, Reading Comprehension, and Writing Ability.

Intercorrelations of the magnitudes cited above tend to suggest a reasonably strong relationship--on a group basis--between skills measured by paper-and-pencil instruments such as the TOEFL and speaking proficiency as judged through the evaluation of active speech production. However, for purposes of making highly reliable statements about the speaking proficiency level of individual test candidates, extrapolation on the basis of group correlational data may be considered a rather questionable procedure: individual variations in the language learning history of given candidates (for example, extensive prior training in one or two specific skill areas with relatively little attention to the other areas) may result in considerable under- or overestimation of speaking proficiency based on the other skills test scores. In situations where it is of considerable practical importance to have a very accurate indication

of the examinee's level of speaking proficiency, there would appear to be no suitable alternative to testing this language skill on a direct and explicit basis.

In recognition of the potential disparity, in any given case, between a candidate's performance on the TOEFL in its regular ("paper-and-pencil") form and the actual level of speaking performance in an active speaking situation--and in recognition also of the need to have a very accurate indication of the candidate's speaking proficiency in order to make informed decisions in such areas as the hiring of non-native English speaking teaching assistants at the undergraduate or graduate level--the TOEFL Research Committee and the TOEFL Policy Council approved in the fall of 1976 the initiation of a research study entitled "An Exploration of Speaking Proficiency Measures in the TOEFL Context." The nature of and results of that study, and its implications for test development within the TOEFL program, are described in the following pages.

General Considerations in Study Design

Both the general scope of the research study and the research strategy employed may be described and analyzed in light of the basic structure and purpose of the TOEFL program and the associated measurement constraints within which it must operate. A major consideration is that any testing instruments used within the program must be amenable to proper and reliable administration at any of the approximately 200 testing centers worldwide. This factor immediately rules out of consideration, at least for operational use within the program, face-to-face interviewing procedures of the type exemplified by the well-known Foreign Service Institute (FSI) technique (Sollenberger, 1978; Wilds, 1975). The FSI interview and similar procedures involving direct conversation with the examinee in the test language require that the interviewer not only be a native or near-native speaker of the language in question but also that he or she be highly trained in the interview procedure. Although test supervisors at the TOEFL centers are highly qualified in the administrative aspects of the testing process (distribution of test booklets, verification of candidate identity, maintenance of proper conditions in the testing room, and so forth) they are in many cases not as highly proficient in the English language as would be required for active interview testing in English. The problem of providing the necessary training in interviewing techniques on a worldwide basis is considered virtually insurmountable on practical and financial grounds if not on a purely theoretical basis.

This major consideration--together with a number of other factors such as, for example, the possibility that differential treatment of interview candidates within or across centers might occur if certain candidates had a friendship or student status with the examiner--ruled out operational use of the FSI interview or similar face-to-face testing techniques in the TOEFL program. However, the very high degree of face/content validity inherent in the direct interviewing procedure as a measure of active speaking skill strongly suggested the use of a direct

proficiency interview as a criterion measure for the study against which other less direct measures could be compared and analyzed.¹

A second factor underlying the planning of the study and the development of experimental testing materials was that these materials could not be designed with a particular candidate first-language in mind, but would need to be appropriate for all TOEFL candidates regardless of mother tongue background. Operationally, this constraint dictated that it would not be possible to make use of a comparative linguistic approach to test content specification (for example, developing test questions exemplifying common English-speaking problems of native Spanish speakers). Instead, both test material development and the design of the associated scoring procedures would have to be based on a "first-language-free" or "first-language-neutral" approach.

A third factor in study planning and test design was the necessity for the developed testing procedures to demonstrate a generalized speaking proficiency on the part of the examinee, rather than proficiency in a particular, delimited language-use situation. In certain other testing contexts--where it is possible to specify minutely and in advance the particular type of employment, area of academic specialization, etc. of the individual(s) to be tested--test content can be explicitly based on the specific language-use situations represented, as suggested, for example, by the situational syllabi developed under the auspices of the Council of Europe (Van Ek, 1975; Coste et al., 1976).

Within the TOEFL program, however, the wide variety of academic or other pursuits represented by the TOEFL population as a whole does not permit "tailor-made" specification of test content and language-use setting but instead requires a more generalized approach, involving the use of test formats and individual test questions that do not require any particular familiarity with given subject matter areas or employment contexts.

Consistent with the major program-related factors identified above, it was determined that the speaking test study would operate within the following basic framework:

(1) Direct (face-to-face) testing techniques would not be considered for operational use within the TOEFL program but would instead be used as criterion measures for the experimental study.

(2) Testing techniques and test formats to be considered for operational use would have to be "automated" to the extent that relatively untrained persons would be able to administer the entire testing process.

¹For further discussion of the basic characteristics of and differences between "direct" (i.e., face-to-face) and "semi-direct" (booklet/tape mediated) testing procedures of the type developed for the present study, see Clark (1979).

(3) Test content could not be developed with a single native-language population in mind but must be appropriate for all TOEFL candidates regardless of mother tongue.

(4) Test content could not be targeted to a particular academic discipline, employment situation, or other specialized language-use situation, but would have to be based on more generalized linguistic contexts.

Additional major decisions relative to the design of the study were as follows:

(1) A reasonably large and heterogeneous group of non-native speakers of English would serve as the experimental population (N of approximately 125-150).

(2) All of the measures used in the study (i.e., a criterion speaking test; experimental speaking test materials; and, for comparison purposes, the regular "paper-and-pencil" TOEFL) would be administered to the experimental group on a concurrent basis to avoid possible confounding effects of improvement in speaking performance in the course of the study itself.

(3) A substantially wider variety of experimental item types would be developed and tried out than would be expected to appear in any operational version of the speaking test. This would provide the opportunity to select, on the basis of the study results, the particular subset of item types found most appropriate for actual use in an operating program.

(4) Prior to the major administration of the experimental materials, a smaller, clinically-oriented administration would be carried out. Clinical administration would provide the opportunity to weed out from further consideration any testing formats or item types observed to have major drawbacks and to further refine the retained formats/item types before they were administered on a larger scale.

(5) Within each format/item type, a sufficient number of items would be administered so that, if the experimental results were encouraging, sufficient items would be available to make up a prototype operational test which could itself undergo additional trial administration and study before being made available for use in the TOEFL program.

THE FSI CRITERION INTERVIEW

The project plan called for, as an initial step, the identification of a test of general speaking proficiency that could serve as a highly face-valid and content-valid measure of communicative ability in real-life speaking situations, and against which the necessarily less direct testing procedures developed for the project could be empirically compared and validated.

The instrument selected for this purpose was the structured conversational interview technique developed by the U.S. Foreign Service Institute in the late 1950's and commonly referred to as the "FSI interview." As designed by the FSI and administered by that institution and increasingly by other groups and organizations both within and outside the U.S. government,² the interview consists of a face-to-face conversation of approximately 15-25 minutes between the examinee and a trained interviewer who is a native or near-native speaker of the test language. The conversation begins at a fairly simple level and becomes increasingly more sophisticated linguistically, as reflected in increased rate of speech on the part of the interviewer, use of more complex structures and more specialized vocabulary, up until the point at which the examinee is no longer able to hold his or her own in conversation at that level. At this point, the level of sophistication of the conversation is reduced somewhat and the examiner spends several minutes exploring the examinee's breadth of command of grammatical structures (for example, ability to use past and future tense forms, conditional constructions, etc.); and extent of active vocabulary, as elicited by questions probing a variety of topical areas including personal and family background, work activities, studies, hobbies and free-time activities, future plans, and so forth. With more proficient examinees, the interviewer will also broach political, social, economic, or other topics requiring very high levels of language use. The interview continues until the examiner is satisfied that the interviewee has fully demonstrated the highest level of speaking proficiency of which he or she is capable.

The interviewee's performance is evaluated on a scale ranging from 0 to 5; 0 denoting no functional proficiency in the language and 5 representing a level of language performance indistinguishable in all respects from that of an educated native speaker of the test language. Each of the numerical levels is accompanied by a brief verbal description of the real-life linguistic situations in which an examinee scoring at that level would be considered capable of functioning effectively. For example, the verbal description associated with level "1" is as follows:

Able to satisfy routine travel needs and minimum courtesy requirements. Can ask and answer questions on topics very familiar to him; within the scope of his very limited language experience can understand simple questions and statements, allowing for slowed speech, repetition or paraphrase; speaking vocabulary inadequate to express anything but the most elementary needs; errors in pronunciation and grammar are frequent, but can be understood by a native speaker used to dealing with foreigners attempting to speak his language. While elementary needs vary considerably

²Organizations currently using the FSI technique include ACTION/Peace Corps, the Bank of Canada, Center for Applied Linguistics, the Central Intelligence Agency, Chula Vista (Calif.) School District, Defense Language Institute, Illinois Bilingual Service Center, National Security Agency, New Brunswick (Canada) Education Department, New Jersey Department of Education, and Texas Education Agency.

from individual to individual, any person at level 1 should be able to order a simple meal, ask for shelter or lodging, ask and give simple directions, make purchases, and tell time.

In addition to the 6 numerical score levels (0-5), a "plus" value is associated with each score level except for 5, resulting in an 11-point scoring range in total: 0, 0+, 1, 1+, 2, 2+, 3, 3+, 4, 4+, 5. The "plus" value is used to denote an examinee performance that substantially exceeds the minimum requirements for a given level but fails to satisfy all of the requirements of the next higher level. Verbal definitions of all five levels are given in Appendix C.

The FSI interview is by no means considered a perfect or ideal measure of communicative proficiency in real-life settings. An obvious shortcoming is that the interview setting cannot directly reproduce the physical surroundings of the great variety of language-use situations (workplace, classroom, clothing store, grocery, friend's house, etc.) in which the examinee would be expected to perform in real life. A second, related, consideration is that the psychological and affective aspects of real-life communication, including motivation and communicative intent of the speakers, status roles of interviewer and interviewee, and a number of other aspects of the real-life situation cannot be precisely duplicated in the interview setting.

In spite of these limitations, the fact that the FSI interview does involve a fairly extensive face-to-face conversation with a native or near-native speaker in a quite realistic setting (by comparison to other possible testing techniques) provides for a high degree of face- and content-validity for this procedure and was considered to justify strong consideration of the FSI interview as a criterion measure against which the necessarily less direct experimental testing formats could be compared and analyzed.³

The degree of scoring reliability that could be anticipated for the FSI-type interview was not well known at the time the study was being planned. Foreign Service Institute staff involved with the interview testing program were confident on the basis of long-term, detailed acquaintance with the procedure that it was being very reliably scored by the FSI testing staff, but no detailed studies of the actual scoring reliability had been conducted as of that time.⁴

³For additional discussion of the development history and measurement rationale of the FSI interview, see Clark (1972), Sollenberger (1978), and Wilds (1975).

⁴A number of such studies have been carried out subsequently to the study conducted for this project, as reported in Clark (1978).

Preliminary FSI Reliability Study

In order to determine the degree of scoring reliability that could be expected for the FSI procedure if it were to be used as the criterion measure for the study, it was decided that an initial investigation of both the inter- and intrarater reliability of the FSI technique--as used under conditions similar to those that would be encountered in the main study--would be carried out before a final decision on the use of the interview would be made.

The basic approach in this preliminary study was to train a number of persons in the FSI interviewing and rating technique; conduct a number of interviews of non-native English speakers using the same procedures that would be used in the main study; score the interviews on the basis of the established FSI rating procedures and scoring scale; and, finally, determine and report the degree of scoring reliability obtained under this testing process.

A total of four persons to be trained as interviewers were identified from among a number of applicants in the Princeton, N.J. area, with final selection being made on the basis of résumé inspection followed by personal interviews. All interviewers finally selected were native speakers of English who also had an excellent technical familiarity with English through various combinations of undergraduate and graduate English study, general and English linguistics courses, and direct experience in ESL teaching.

The four interviewers underwent an intensive two-day training session at ETS in which project staff explained in detail the interviewing process and interview scoring procedures and conducted live interviews for group discussion and critique. During the afternoon and evening of the two training days, each participant listened to a series of 15 "training tapes" consisting of a series of sample interviews ranging from 0+ to 4+. The two final steps in the training process were to have each participant conduct one or more live interviews for evaluation and critique by the ETS staff and the other participants; and to listen to and rate individually a second series of 15 tape-recorded interviews, presented in random order, for which the official interview scores were not known by the rater in advance. On the basis of all of the training activities, and including a consideration of interviewing performance and accuracy of rating of the testing tapes, each of the four participants was considered qualified for both interviewing and rating, based on the general standards used by ETS in the training of interviewers/raters for other testing programs.

About three and a half weeks after the initial training process, the four approved interviewer/raters and the senior author conducted a three-day interviewing session with a group of undergraduate and graduate ESL students enrolled in the American Language Program (ALP) at Columbia University. A total of 86 students participated: 49 men and 37 women ranging in age from 17 to 61 (S.D.=8.57). A total of 26 different native languages were represented by the group.

Over the three-day period, students were scheduled to appear for the interview testing at 30-minute intervals. On arrival at the testing site, each participant was asked to complete a short questionnaire giving basic background information. This questionnaire also asked the participant to indicate his or her instructional (course placement) level at the ALP as well as to provide a self-rating of speaking proficiency by reading over the FSI verbal descriptions for levels 0-5 and placing a check mark opposite the rating which the participant considered to most accurately reflect his or her current level of speaking ability in English. Questionnaire data were not made available to the interviewer/raters until the interview rating process had been completed to avoid any possible influence on the interview scores assigned.

As an additional aspect of the study, each of the student participants was asked to take a considerably briefer (5-6 minute) interview as well as the regular (15-20 minute) interview. The purpose of this was to determine whether the shorter interview could be scored with a reasonably high degree of reliability. If this were the case, it might be possible for certain testing purposes (though not necessarily for the speaking test development study itself, in which the highest possible scoring reliability was being sought) to make use of an appreciably shorter--and by the same token more economical of time and human facilities--interviewing process.

Interviewing was counterbalanced in such a way that approximately half the participants received the short interview first and half the long. For a given interviewee, the long and short interviews were given by different interviewers to avoid any carry-over or "halo" effect on the ratings made. Each interviewer conducted long and short interviews for approximately the same total amount of time. Three interviewers began the testing session with long interviews and two with short interviews to counterbalance any effects of short and long interview sequencing across interviewers.

For both the long and short interviews, ratings were given on-the-spot by the person conducting the interview. In addition, each interview was tape recorded for later evaluation, using a cassette recorder with automatic recording level control and two small lapel microphones, one for the interviewer and one for the interviewee.

Approximately ten days after the on-site interviewing session at the ALP, each of the five interviewer/raters listened to and rated all of the tape-recorded long and short interviews for each of the student participants. For this rerating, the interview cassettes were sequenced so that, for each rater, groups of approximately 15 short interviews followed by 15 long interviews were presented (or vice versa). In no case were the long and short interviews of a given examinee scored "back-to-back"; all were separated by at least 15 intervening interviews. Conversations with the raters on conclusion of the rating process indicated that the raters did not remember any of the scores initially assigned to the examinees, with the exception of one or two examinees at the highest or lowest extremes of the scoring scale, whose initial scores were recalled because of their

divergence from most of the other ratings given. However, because of the great number of interview tapes evaluated in total, it is considered that for all practical purposes the rerating of interviews was not influenced by knowledge of earlier ratings.

Intrarater Reliability

Each of the five interviewers had initially interviewed approximately 17 students by means of the long interview and 17 additional (different) students using the short interview. Thus, during the rerating process--in which each interviewer/rater listened to all of the interviews conducted over the three-day period--each of his or her "own" interviews, both long and short, were also rerated. The intrarater reliability information thus obtained for each of the five raters is shown in Table 1 below.

Table 1
Intrarater Reliabilities for Long and Short Interviews

Long Interview			Short Interview		
Rater	r	N	Rater	r	N
A	.907	17	A	.837	17
B	.868	17	B	.904	14
C	.947	19	C	.853	15
D	.771	17	D	.740	18
E	.840	15	E	.751	11

Intrarater reliabilities of the long interviews ranged from .771 to .947, with an average reliability of .867. Reliabilities for the short interviews were only slightly lower--ranging from .740 to .904, with an average reliability of .817.

Interrater Reliability

Interrater reliability coefficients for all possible pairs of raters are shown in Table 2. Because of certain administrative problems in distributing the short interview recordings to all raters, it was not possible in some cases for all five raters to score a particular short interview. Interviews for which even a single rating was missing among the five judges were excluded from the interrater analysis, resulting in a total of 68 short interviews for which complete scoring information for all five raters was available. For the long interviews, there were 86 complete scoring records.

Table 2

Interrater Reliabilities for Long and Short Interviews

Long Interviews (N=86)

<u>Rater</u>	A	B	C	D	E	
A	1.000					
B	.840	1.000				
C	.602	.705	1.000			
D	.780	.788	.712	1.000		
E	.814	.804	.593	.711	1.000	Mean $\bar{r}$ = .735

Short Interviews (N=68)

<u>Rater</u>	A	B	C	D	E	
A	1.000					
B	.857	1.000				
C	.778	.741	1.000			
D	.771	.767	.744	1.000		
E	.752	.782	.679	.709	1.000	Mean $\bar{r}$ = .758

Interrater reliability, for both the long and short interview formats, appeared to be somewhat lower than had been anticipated, although in any case reasonably high for a non-objectively scored test of the type represented by the interview. These results suggested that if the FSI interview technique were indeed to be used as the criterion measure in the main study, it would be desirable to increase the scoring reliability by having the interviews rated twice and combining these ratings into a single total rating to serve as the criterion score for a given examinee.

Correlations with External Criteria

The project staff's consideration of the appropriateness of using the FSI interview as the criterion speaking test measure was also based on an examination of correspondences between the interview data and other available "external" data, which included: (1) the instructional level at the ALP to which the examinee was presently assigned (and presumed to reflect the examinee's general level of proficiency in English); (2) the examinee's score on a multiple-choice placement test developed at the ALP and consisting of 60 tape-recorded listening comprehension questions and a 120-question section dealing with English grammar and vocabulary; and (3) the examinee's self-rating of speaking proficiency based on the 0-5 FSI scale. Correlations of these additional variables with both the long and short interview scores are shown in Table 3 below.

Table 3

Intercorrelations of Long and Short Interview Scores
and Other Indices of Language Competence

	(1)	(2)	(3)	(4)	(5)
1. Instructional Level at ALP	1.000	.590	.558	.610	.551
2. ALP Placement Test Score	.590	1.000	.348	.570	.707
3. Self-Rating of Speaking Proficiency	.558	.348	1.000	.479	.430
4. Long Interview Score	.610	.570	.479	1.000	.696
5. Short Interview Score	.551	.707	.430	.696	1.000

All of the correlations in Table 3 are significant at $p = .01$ or beyond. Of the three external variables, examinee self-ratings had the lowest correlations with both long and short interview scores. This may be attributable to some extent to two factors: first, the fact that provision was not made for the interviewees to indicate "plus" values (which would have resulted in a more extensive and possibly more precise rating scale); and second, the probability that many of the examinees, especially those at the lower proficiency levels, were not able to read the rather sophisticated verbal definitions with full comprehension.

Taking into account these and other factors that may have operated to reduce the magnitude of the self-rating/interview score correlations, the observed figures are considered to demonstrate a reasonably substantial relationship between the interview ratings received by the examinees and their own perceived levels of speaking proficiency. To the extent that the examinees' own appraisals of their speaking ability can be considered a valid indication of the "true situation" in this regard, the obtained self-rating/interview score correlations would tend to provide additional supporting evidence of the validity of the interview process as an appropriate measure of speaking skill.

Both the instructional level at the ALP and the ALP placement test scores show reasonably high (.551 to .707) correlations with the long and short interview scores. Although variables other than absolute speaking proficiency per se enter into both the placement test and instructional level data, these two additional measures--especially the instructional level designation, which is largely based on placement test scores but also reflects subsequent changes in classroom assignments based on teachers' observations of the examinees' language performance in the classroom--were considered to constitute additional supporting data in this regard.

Considerations in Adoption of FSI Criterion

In making the final decision on use of the FSI interview as the criterion measure for the main study, project staff considered available data from the initial scoring study as well as a number of other factors. First, although the observed scoring reliabilities for the interview--as based on scoring by a single rater--were not extremely high, they were considered reasonable for non-objectively scored tests of this general type. In this regard, it was considered possible to increase the scoring reliability in the main part of the study by having each interview scored by two raters rather than one.

Second, examinee self-ratings and other measures presumed to relate positively to speaking proficiency were found to correlate at a reasonably high level with the FSI interview scores, providing additional external validation of this technique.

Third, the high degree of face validity of the conversation-based interviewing technique, together with a scoring scale based (for the most part) on statements of real-life language use capabilities on the part of the examinee, were considered to provide measurement and interpretation advantages for the study which were not closely met in any other available testing procedure.

Fourth and finally, the widespread use of and familiarity with the FSI procedure as a measure of general speaking proficiency in other research studies and in a variety of practical applications meant that examinee scores expressed on this scale--and, by the same token, scores on any eventual TOEFL-program speaking test referencing the corresponding FSI scale levels--would be immediately meaningful to a number of outside audiences having an interest in the research study or its developmental outcomes.

DEVELOPMENT OF INITIAL EXPERIMENTAL TEST FORMS

As previously discussed, the speaking test procedures considered appropriate for use within the operational constraints of the TOEFL program necessarily involved the "semi-direct" testing of candidate performance by means of test questions/stimuli presented in a printed test booklet or tape recording, rather than through "face-to-face" conversation with a live interlocutor. However, even within the limitations imposed by a tape/booklet testing format, there was a very considerable range of possible elicitation procedures and question types that could be considered for use, as suggested in the numerous examples and associated discussions available in the language testing literature.⁵

ETS project staff--following detailed review of these sources, examination of a variety of question types used in previously developed

⁵Especially relevant sources include Lado (1961), Valette (1967), and Clark (1972, 1975).

"semi-direct" speaking tests,⁶ and a series of meetings aimed at the identification and discussion of other potentially useful testing procedures--identified a total of 11 separate formats/question types which it was felt should be incorporated into the "first run" of experimental test materials for the study. These included item types eliciting highly discrete language elements on the part of the examinee (for example, particular lexical items or syntactical/morphological patterns) as well as more generally oriented formats in which the examinee would have a considerably greater degree of freedom in response.

Other factors being equal, the more global item types were to be preferred for eventual operational use, in keeping with the basic function of the test as a measure of generalized speaking proficiency. However, it was also recognized that it would probably be necessary to include a certain number of somewhat more highly structured item types in the final test in order to maintain a satisfactorily high degree of scoring reliability for the test as a whole. By exploring and comparing the performance of a considerably greater number of item types than would ultimately appear in the test, it would be possible for the project staff to select a "most-potentially-useful" subset of item types for eventual operational use. This selection process would involve the simultaneous consideration of a number of factors, including the degree of face validity of each item type as a measure of general speaking proficiency, observed correlation with the FSI proficiency interview, scoring reliability, ease and straightforwardness of administration, and the linguistic content and statistical performance of that particular item type in comparison with (and in the context of) other item types with which it might appear in a final operational test.

Each of the 11 item type formats proposed for inclusion in the initial experimental version of the speaking test is described in some detail below, together with a discussion of the rationale for its inclusion, perceived strengths and drawbacks of the procedure, and other relevant observations.

Autobiographical Questions (AUTOBIOG)

The examinee hears short, biographically oriented questions on the test tape and responds during the pause following each question. No printed material appears in the test booklet.

Example:

(Master Voice on Test Tape)	(Examinee Response)
May I have your name, please?	My name is....
How many brothers and sisters do you have?	I have two brothers and one sister.
What are your plans for the coming year?	I am planning to visit some friends in California....

⁶For example, the speaking test sections of the MLA-Cooperative Foreign Language Tests and MLA Foreign Language Proficiency Tests for Teachers and Advanced Students batteries (ETS, 1963; 1961)

The face/content validity of this question type is fairly high in that similar questions would probably be asked of the examinee in such real-life situations as academic advising sessions or employment interviews. This format is quite economical of testing time since instructions to the examinee can be very brief (essentially, "Please answer the following questions about yourself"), and the questions themselves are intrinsically short and succinct.

With separate scoring of each question/answer unit, scoring reliability for this section of the test was expected to be quite high. Potential disadvantages included the possibility that most examinees would have "memorized," or at least be very well versed in, the kinds of question-answer situations represented by these rather basic autobiographical items and that as a result there would be relatively little score variation across examinees. The scope of questioning possible in this item type is also fairly limited, since the questions must be potentially answerable by any examinee, regardless of his or her personal background or present or future plans. (For example, questions about past or prospective courses of study would not be appropriate for non-academic examinees taking the test for employment or other non-school related purposes.)

Spoken Vocabulary - Nouns (SV-N)

In this question type, the examinee looks at simple numbered line drawings of objects or persons reproduced in the test booklet. For each, the examinee is asked to "say the name of that object" or "give the name of that person's job," during the short pause provided for consideration of that picture.

Example:

(Printed in Test Booklet)



(Examinee Response)

[It's a] fork.



[She's a] nurse.

Two major advantages of this item type are great economy of testing time (in that a large number of separate questions can be presented in a relatively short period); and, as a corollary, the potentially very high scoring reliability for this section as a result of the many separate "scorable items" presented and the simple dichotomous (right-wrong) scoring decision to be made in each case.

At least three potential disadvantages or cautions regarding items of this type may be identified. First is the matter of cultural fairness or "cultural neutrality" of the objects/persons depicted. Any objects or persons represented that would not be highly familiar to all examinees as a result of general experience in their home countries would be considered inappropriate for inclusion in the test (for example, a picture of a snow plow). Second, the drawings themselves would have to be highly unambiguous as to the person or object represented.

Associated with the two preceding considerations was the concern that it might not be possible to present any but the most simple and straightforward kinds of vocabulary items in this section--items that most examinees could be expected to know as a matter of course even at the lowest levels of general proficiency. Although this was considered a very possible outcome, it was decided to include this item type, at least for purposes of experimental tryout, in the initial version of the test.

Spoken Vocabulary - Verbs (SV-V)

In this item type, the examinee sees simple line drawings of persons engaged in a specific action and is asked to "say what they are doing." A short pause is provided for response to each picture.

Example:

(Printed in Test Booklet)



(Examinee Response)

[He's/The man is] running.



[She's/The woman is] reading
[a book].

Advantages to and potential drawbacks of this item type are essentially the same as for the noun-vocabulary items. A further limitation in the use of pictured actions is that the examinee's command of "higher-order" vocabulary (i.e., more sophisticated or more precise expressions as distinguished from more basic or "generic" expressions) cannot readily be elicited through pictorial means. For example, a drawing such as the one shown below, although intended by the test developer to portray "slipping," could quite reasonably be interpreted (and responded to) as the more general "falling" by native as well as non-native English speakers.



Reading Aloud (RD ALOUD)

The examinee sees a printed paragraph of about 125 words and is given one minute to read the paragraph silently. The examinee then reads the paragraph aloud with attention to pronunciation, intonation, and style/tone of delivery.

Example:

(Printed in Test Booklet)

Despite the decrease in size--and, some would say, quality--of our cultural world, there still remain strong differences between the usual British and American writing styles. The question is, how do you get your message across? English prose conveys its most novel ideas as if they were timeless truths, while American writing exaggerates: if you believe half of what is said, that's enough. The former uses understatement; the latter, overstatement. There are also disadvantages to each characteristic approach. Readers who are used to being screamed at may not listen when someone chooses to whisper politely. At the same time, the individual who is used to a quiet manner may reject a series of loud imperatives.

Reading aloud from a printed text, although of considerably less frequent occurrence than face-to-face conversation, is an occasional component of language use in real-life situations (e.g., reading a report in class, making a presentation at a group meeting) and as such is considered to have a useful degree of face/content validity in the speaking test setting. The reading aloud of a single specified text by all examinees (as contrasted to the use of more "open-ended" kinds of test exercises) would also be expected to permit a fairly high degree of scoring reliability.

On the negative side, testing efficiency is reduced by the rather lengthy "dead time" during which the examinee reads the passage silently--time that could be devoted to active responding in some other item type format.

Sentence Repetition (SENT REP)

The examinee hears a series of English sentences of increasing length and semantic complexity and is asked to repeat each sentence verbatim during the pauses provided. Sentences are not shown in the test booklet.

Examples:

(Heard on Test Tape; To Be Repeated Exactly)

Tom worked late yesterday.
He hasn't had a chance to look at the mail yet.
We left the museum and went to a nearby restaurant for a cup of coffee.

The sentence repetition technique has rather low face/content validity in that verbatim repetition of spoken stimuli is not generally called for in real-life situations, except for such minor and infrequent situations as repeating a just-heard address or telephone number to verify correctness. However, this item type was considered very much worth investigation on both psycholinguistic and psychometric grounds.

From the psycholinguistic standpoint, examinee performance on the sentence repetition task could be considered to involve the process of "chunking." The working hypothesis in this regard is that examinees with a higher level of English proficiency would be able to store in and reproduce from short-term memory reasonably large "chunks" of information consisting of whole phrases, clausal units, or other multiple-word segments of the stimulus sentence. Less proficient examinees, on the other hand, would find it necessary to process as chunks much smaller segments of the stimulus sentence or even individual words, leading rapidly to an overflow of short-term memory and the consequent inability to retain and reproduce substantial portions of the stimulus sentence.⁷

From a psychometric standpoint, the sentence repetition format was considered to offer the possibility of high scoring reliability in that the scorers would be easily able to distinguish between correct (i.e., verbatim) and incorrect (not-verbatim) repetitions of the stimulus sentences. These two considerations of underlying psycholinguistic theory and potentially high scoring reliability were considered sufficient justification for including the sentence repetition item type among the techniques to be tried out in the experimental administrations.

⁷(For further explanation and discussion of the "chunking" hypothesis see, for example, Clark and Clark, 1977.)

Combining Sentence Elements (SENT ELEM)

For each question in this section, the examinee sees three to four different "sentence parts," each separated by slashes, and is asked to combine these parts into a complete, meaningful sentence by adding appropriate prepositions or other connectives and/or by making suitable morphological changes in the root words provided. A short time is allotted for the examinee to study the sentence parts and to say aloud a complete sentence using these parts.

Example:

(Printed in Test Booklet)

Library/stay open/midnight

(Examinee Response)

The library stays open until midnight./The library stayed open until midnight [or other grammatically acceptable reply using all three elements].

This item type may be considered to occupy an intermediate position between highly open-ended and completely fixed response modes. The examinee is constrained to produce a response which incorporates each of the sentence elements given, but has some flexibility in determining the final form of the sentence produced. A moderately high level of scoring reliability was anticipated, and it was felt that an item format of this type would provide a useful bridge between the more highly controlled formats in the earlier part of the test and the freer-response formats to be encountered later on. Although face validity of this item type is quite low (in that slash-separated sentence parts would not under ordinary circumstances be seen or require dealing with in real life), the process followed by the examinee in responding to such stimuli--specifically, the "fleshing out" of core words or concepts in a morphologically and syntactically acceptable form--was felt to parallel quite closely the sentence-generation procedures followed by native speakers in real communicative contexts.

Grammatical Patterns (GRAM PAT)

The examinee sees in the test booklet a model pattern for a tense or number change or other morphological/syntactical manipulation, accompanied by both printed and spoken explanations. The examinee is then asked to carry out similar manipulations for other items based on the same pattern.

Example:

(Printed in Test Booklet)

[Together with necessary preliminary instructions]

Model: (You see:) Many people are afraid _____ dogs.

(You say:) Many people are afraid of dogs.

(Printed in Test Booklet)

(Examinee Response)

Tom is interested _____ new science exhibit.

Tom is interested in the new science exhibit.

They're waiting _____ the show to start.

They're waiting for the show to start.

Of all the item types proposed for inclusion in the experimental test form, the grammatical patterns type was considered the most open to question. A major concern focused on potentially undesirable influences on instructional practices that might occur if examinees preparing for the test were to concentrate unduly on grammatical "pattern drills" to the detriment of other kinds of speaking practice. A second concern involved the problem of adequate content sampling: since each separate "pattern" to be tested in this section would require a separate model and related instructions, it was felt that no more than three different patterns could be included without seriously overweighting this part of the test with respect to the other test sections. In terms of the great number of grammatical patterns that might be proposed as candidates for testing, the inclusion of only three such patterns in a given test would not begin to approach an adequate sampling of this domain.

Despite these rather serious reservations, it was decided to include a pattern practice section in the test materials, primarily for experimental and comparative purposes rather than as a serious candidate for inclusion in the operational test.

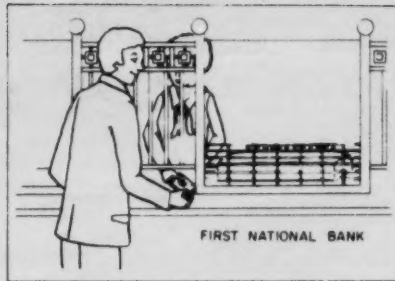
Description of Picture Sequence (PICT SEQ)

In this section, the examinee sees a panel of six line drawings depicting a chronologically ordered sequence of events and is asked to "tell the story that the pictures tell," using the past tense. After studying the pictures silently for 30 seconds, the examinee is given one minute to describe the pictured events.

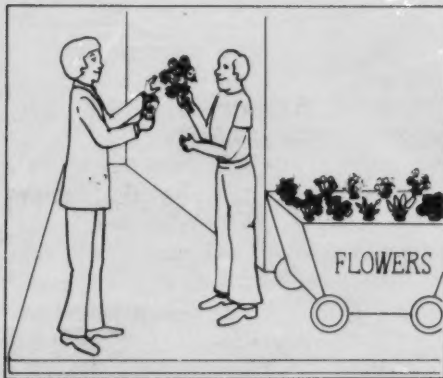
Example:
(Printed in Test Booklet)



1



2



3



4



5



6

(Examinee Response)

[A complete description would be expected to include the concepts of "withdrew money/took money out of the bank," "bought flowers," "gave the flowers to his girlfriend/wife," "went to see a movie/play," and "drank some coffee/had something to drink, etc. after the movie/play."]

This format is considered to reflect quite closely real-life situations in which the speaker "tells a story" or engages in some other relatively lengthy narration of events. It uses available testing time very efficiently in that, following brief instructions, the examinee provides an extensive sample of his or her speech, completely uninterrupted by test directions, spoken question numbers, or other intrusive elements.

Although quite open-ended (more so than any of the preceding question types), the picture sequence format defines the general topic to be described and sets the basic parameters within which all examinees are expected to respond. Provided that the stimulus pictures themselves are unambiguous and culturally neutral, they constitute a very straightforward means of eliciting the desired responses and, of course, do not require any proficiency in reading English on the part of the examinee.

Multiple Questions on Single Picture (MULT Q)

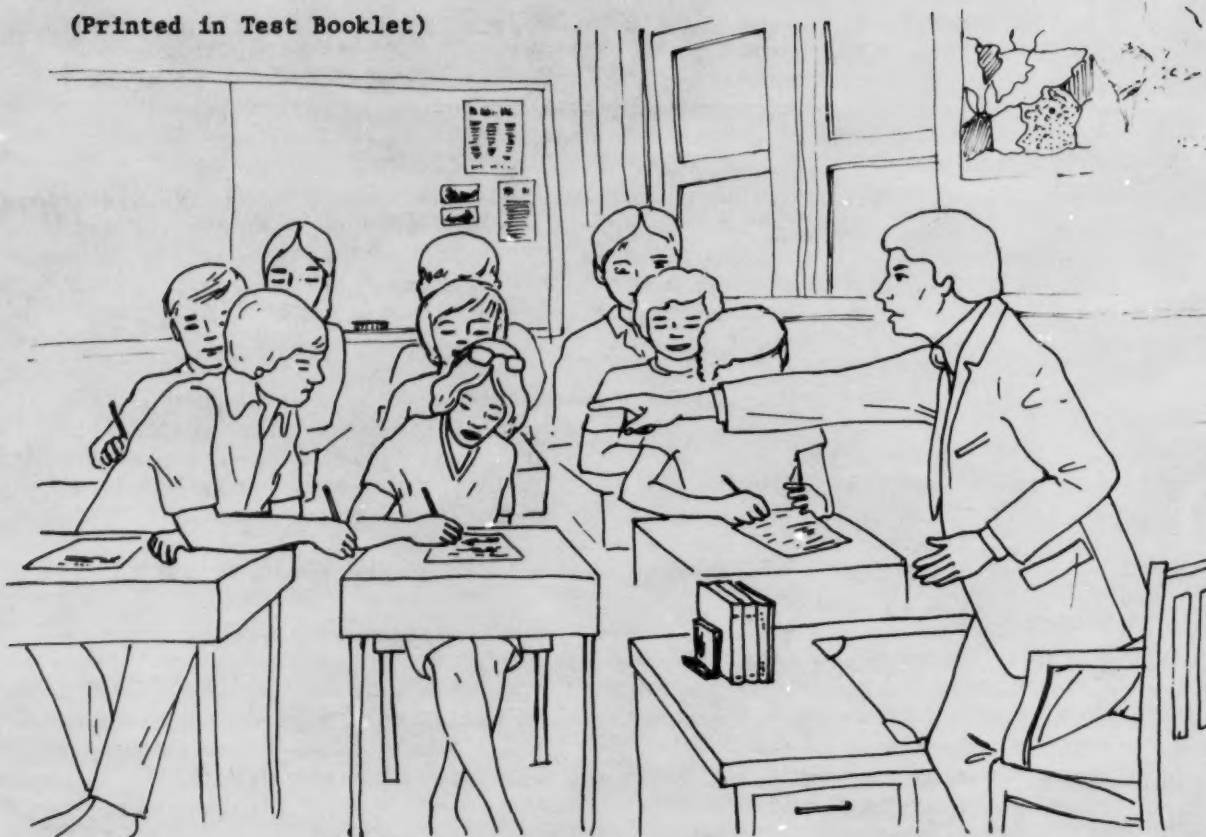
The examinee studies a single line drawing for 30 seconds and then answers a series of spoken questions about the content of the picture, events leading up to the situation represented by the picture, probable thoughts or attitudes of the persons in the picture and future events that might take place as a result of the situation depicted. The questions themselves are not reproduced in the test booklet.

This format is similar in many respects to the picture sequence approach but is considered somewhat less straightforward and natural because the master voice must interpose each new question. By the same token, the examinee is permitted to take a less active role in the speaking process in that he or she has only to answer the specific questions posed and then passively wait for the next question, rather than integrate a number of concepts into a continuous narrative, as is the case with the picture sequence format.

Both the picture sequence and the single-picture formats were considered to be less reliably scorable than the more discrete item types presented earlier in the test, but it was felt that the increased realism and opportunity to demonstrate fluency of the picture-based formats would, on a total test basis, more than offset the presumed lower scoring reliability for these particular sections.

Example:

(Printed in Test Booklet)



(Questions Heard by Examinee)

Where is this scene taking place?

What is the teacher doing?

What is the teacher thinking?

What will probably happen to the boy?

(Typical Response)

[In] a classroom.

Scolding/yelling at the boy.

[The teacher thinks] the boy
is cheating/copying
from the girl, etc.

He will be thrown out of the
class/have to stay after
school/get a zero on the
test, etc.

Simulated Telephone Conversation (TEL CONV)

In this section, the examinee is asked to carry out an imaginary telephone conversation with a voice on the test tape. Before beginning the conversation, the examinee is given time to read (silently) a description of the background and content of the conversation, including a brief synopsis of the questions or statements to be made by the voice on the tape and the anticipated replies. Then, a telephone bell is heard and the conversation begins. During the conversation, the examinee sees in the test booklet a series of one-line "telegraphic" reminders of the replies to be given in each instance.

Example:

(Synopsis Printed in Test Booklet)

You are a student at Northern State University. The spring vacation period begins next week, and you would like to go on a special bus tour to Niagara Falls which is being arranged by the student travel service. You call the travel service and ask about the tour. Your first question is when the tour begins, and the clerk replies that the tour bus will leave next Monday morning. She asks if you have any other questions and you ask her how much the trip will cost and if meals and a place to sleep are included. After she answers, you ask her where you will be staying in Niagara Falls and say that you would rather stay in a home than in a hotel. She replies that you will be staying with an American family and that you had better make your reservation quickly if you want to go. You say that you will come to the travel service office that afternoon and thank her for her help.

(Telegraphic Cues Printed in Test Booklet)

BUS TOUR TO NIAGARA FALLS. WHEN?

COST? WHAT INCLUDED?

WHERE STAY? LIKE TO STAY IN A HOME

COME TO OFFICE THIS AFTERNOON. THANK YOU

(Examinee Response)

[Varies; acceptable response will cover all the points outlined in cues.]

The telephone conversation format enjoys a high degree of face validity in that communication by phone is a pervasive feature of real-life language use. However, a number of technical considerations make effective "tape surrogate" testing of telephone (or other conversational) situations

rather difficult. A major drawback is the inability of this procedure to incorporate the give-and-take of natural conversation: since the taped questions/replies are fixed in content (as well as in timing), an anomalous reply by the examinee or the taking of too much or too little time in responding cannot be adjusted for and may lead to confusion and loss of composure on the part of the examinee. In addition, the amount of time required to present the necessary instructions, including the silent reading of the conversational summary by the examinee, is relatively great in comparison to the amount of examinee speech elicited in this section. Despite these shortcomings, it was decided to experiment with the simulated telephone conversation format in view of its strengths from a face validity standpoint.

Persuasive Speech (PERSUAS)

In this section, the examinee is placed in a role-playing situation in which he or she is required to engage in a "persuasive" interchange with an authority figure represented by the voice on the test tape. The nature of the task is outlined in a brief printed paragraph which the examinee reads before beginning the conversation. During the conversation, the examinee's recollection of the points to be communicated is aided by a panel of four line drawings depicting the interlocutor face-to-face with the authority figure in the situation described. In addition, cartoon-like balloons above the interlocutor's head provide further telegraphic reminders of statements or replies to be made at those points in the conversation.

Example:

(Synopsis Printed in Test Booklet)

You are sitting in a cafeteria eating lunch. You become very dissatisfied with the food and you complain to the manager. First, the coffee you bought is so weak you can hardly taste it. Since coffee is expensive, you want your money's worth. Also, your soup tastes awful. Finally, you have found a piece of glass in your salad. You are very angry about the situation and you demand your money back.

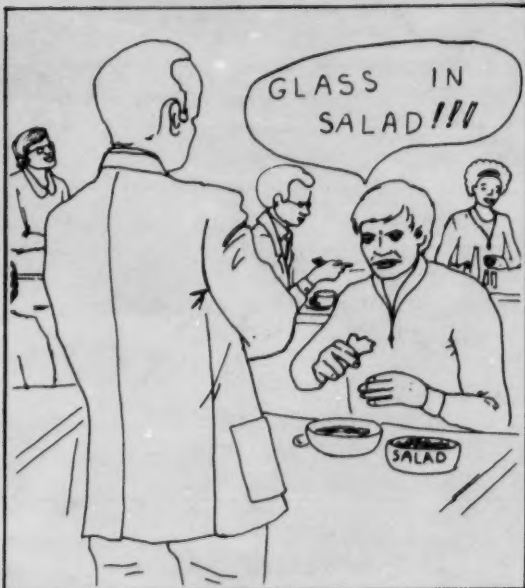
(Picture Cues Printed in Test Booklet)



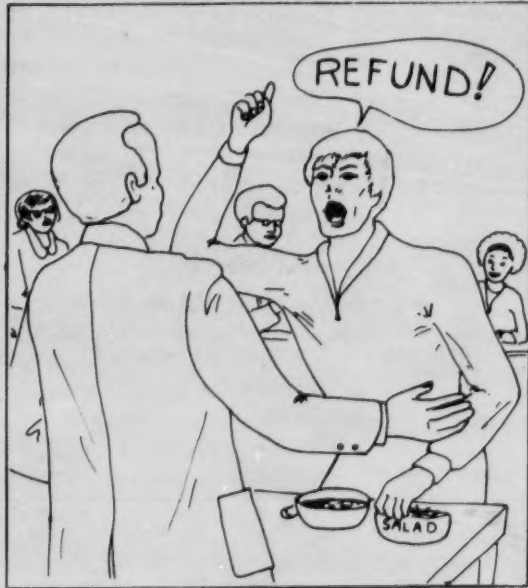
1



2



3



4

(Examinee Response)

[Varies; acceptable response will cover all the points outlined in picture cues.]

This testing format is intended to evaluate speaking proficiency in situations requiring affectively conditioned responses, and involving selection of appropriate register by the examinee as well as effective use of intonation and emphasis in "getting the points across."

The persuasive speech situation shares with the telephone conversation a high degree of relevance to actual communicative situations. Shared disadvantages are the relatively complicated instructions required to set the stage for the examinee's performance and the necessarily fixed pace and sequencing of the taped exchanges.

Final Content of Initial Test

The initial version of the experimental speaking test, prior to any modifications on the basis of the clinical administration, contained each of the item types described above. The sequencing of these item types within the test, the number of individual stimulus questions (where applicable) for each item type, and the actual running time for each item type section are shown in Table 4. For the clinical testing, a sufficient number of questions were prepared to produce two separate test forms, each with a separate test booklet and tape recording.

DEVELOPMENT OF SCORING PROCEDURES

The specific scoring procedures (and associated scoring key) for each question type within the test were considered subject to experimentation and modification in the course of the study in the same way as the testing procedures themselves. As with the latter, there were several basic parameters, based on administrative considerations within the TOEFL program, which governed the design of the scoring procedures to some extent.

First, it was considered necessary to design a scoring system which would allow the persons evaluating the student response tapes to carry out their tasks as efficiently and economically as possible, consistent, of course, with maintaining adequate scoring validity and reliability. A prime consideration was to avoid, insofar as possible, the need for raters to stop the student response tape in the course of scoring or to listen more than once to any given response.

The most desirable situation in this regard would be to have the student response tape running continuously, with the rater making all necessary scoring decisions and marking the scoring form on a real-time basis. A somewhat less optimum approach would be to have the rater stop the tape periodically if necessary to "catch up" with the scoring. Least desirable would be a scoring procedure requiring the rater to relisten frequently to previous portions of the tape (as might be required, for example, in a scoring scheme requiring complex judgments about each of

Table 4

Item Type Content of Initial Test Forms

<u>Section</u>	<u>Content of Section</u>	<u>Directions</u>	<u>Running Time*</u>		<u>Total</u>
			<u>Items</u>		
AUTOBIOG	7 questions	22 sec.	105 sec.		127 sec.
SV-N	32 pictured nouns	50	275		325
SV-V	15 pictured verbs	46	165		211
RD ALOUD	single reading passage	97	64		161
SENT REP	10 questions for repetition	51	126		177
SENT ELEM	7 elements for expansion	83	82		165
GRAM PAT	3 patterns; 5 items per pattern	200	154		354
PICT SEQ	panel of 6 pictures for continuous narrative	63	62		125
MULT Q	6 questions on single picture	42	97		139
TEL CONV	telephone conversation with 6 anticipated replies by examinee	143	92		235
PERSUAS	"face-to-face" conversation cued by 4 pictures; examinee response anticipated for each	100	49		149

*Times shown are for Form 1 of the test; Form 2 timing is essentially identical.

several aspects of the examinee's response). In this case, not only is active scoring time increased but unproductive rewinding time is added to the total period required for scoring.

A second consideration was that for form-to-form consistency and ease of interpretation and use by the scorers, the scoring key should remain uniform across different versions of the test. This would rule out, for example, a scoring key for the reading-aloud section based on the pronunciation of individual words, consonant clusters, etc. that happened to appear in that particular text but that would not be paralleled in the reading texts for other forms of the test.

Third was the general need to keep the scoring procedures as simple and as straightforward as possible and to avoid the necessity for large conceptual shifts on the part of the raters in moving from one question type to another. This implied the use of very similar or identical scoring keys for more than one question type whenever this could be done without prejudice to the measurement intent of that question type.

The full set of scoring keys, as developed for the initial administration of the two experimental test forms, is reproduced in Appendix A. In most cases, rating scales ranged from 0-3 or 0-4, with 0-1 and 0-2 utilized for the more highly discrete item types. In the latter case, items were used in sufficient quantity within a type to yield quasi-continuous item type totals when items were summed.

CLINICAL ADMINISTRATION OF INITIAL TEST FORMS

For the clinical administration and evaluation of the experimental item types as previously described and summarized in Table 4, a group of 12 examinees was convened from Princeton, N.J. and the surrounding area. The participants were all non-native speakers of English, with the following languages as the mother tongue: Chinese (2), Hebrew (3), Japanese (2), Persian (Farsi) (2), Spanish (2), and Turkish (1). Length of stay in the United States ranged from two months to three years. Prior to taking the two forms of the experimental speaking test, each participant was administered a current TOEFL examination and a face-to-face proficiency interview which was later scored by two separate raters using the standard FSI rating procedure.

Across the 12 examinees, TOEFL scores ranged from 323 to 627 and FSI scores--averaged for the two raters--from 1.35 to 4.00. ("Plus" values were coded as 0.7; i.e., level "1" = 1.0, "1+" = 1.7, etc.)

Administration of the experimental test forms was carried out on an individual basis for each examinee, using two cassette tape recorders--one to play the test tape and a second to record the examinee's responses to the stimuli presented by the test tape and/or test booklet.

Each examinee was tested in a quiet, separate room. A project staff member was continuously present in each case to give basic directions on equipment use and, once the test had begun, to observe and unobtrusively take notes on several different aspects of the testing process, including:

- (1) apparent clarity or lack of clarity of the test directions;
- (2) suitability of the pause lengths provided on the tape for the examinees' responses;
- (3) possible flaws in individual items, as indicated by examinee confusion, unanticipated responses, and so forth;
- (4) affective reactions of the examinees to the testing process, both for the test as a whole and for individual item types.

All candidates took both Form 1 and Form 2 of the experimental speaking test, some on the same day (following a rest break) and some on two separate days. Despite the total running time for each of the two forms (approximately 40 minutes each), the great majority of examinees expressed or evidenced no notable degree of testing fatigue. The sole exception was the examinee who had scored lowest on both the TOEFL and the FSI interview (323 and 1.35, respectively); this examinee found the two speaking tests quite enervating and required frequent intervention and support by the observing staff member in the course of the testing.

Since the primary goal of the clinical administration was to obtain observational information on the overall testing procedure and on examinee approaches to the various item types presented (as opposed to developing extensive statistical data of the type that would be gathered in the later large-group administration), a full-scale scoring procedure was not carried out for each of the 24 tests. However, a total of 14 tests were fully scored by project staff members, with some additional spot scoring of individual test sections. From this information, together with detailed observational notes on the test administration and examinee reactions, the following judgments were made concerning each of the experimental item types. Where applicable, recommended changes in the scoring procedure are also shown.

Autobiographical Questions (AUTOBIOG)

The examinees appeared to respond readily to the questions asked. Since single-word answers--scored "0" in the original scoring key--were often given by the examinees as correct and complete answers to a particular question (for example, "And where were you born" answered as "Japan"), it was decided that the scoring key would be changed to allow credit for short but accurate and responsive answers, as follows:

- 0 = no response or fails to communicate
- 1 = communicates, but with grammatical error(s)
- 2 = communicates, no grammatical error.

Spoken Vocabulary - Nouns (SV-N)

Examinee responses in this section pointed to a number of ambiguous pictures requiring alteration or replacement (for example, a picture intended to represent a "suitcase" also elicited "briefcase"). Objects represented by several other pictures (e.g., "window") were answered correctly with such high frequency that it appeared they would have little if any discriminative power in the test.

From the positive standpoint, scoring of this section was very efficient and highly objective by comparison to other item types. Based on the clinical results, it was decided that the noun vocabulary section would be retained for the full-scale administration. However, several ambiguous pictures would be revised or replaced prior to the full-scale administration and extremely easy items would also be replaced by somewhat more difficult nouns.

Spoken Vocabulary - Verbs (SV-V)

This item type exhibited many of the same visual ambiguity problems as the noun vocabulary section (for example, a kitten "scratching" a post was considered by some examinees to be "climbing") as well as a number of inordinately easy questions (for example, a picture of a man "drinking" from a glass). In view also of the previously-discussed problem associated with pictured actions--the difficulty of unambiguously representing more complex or more highly qualified actions (for example, "slipping" vs. the more general "falling")--it was decided that the limitations of this item type were sufficient to rule out its inclusion in the large-scale administration.

Reading Aloud (RD ALOUD)

This section appeared to operate very smoothly for the most part. Rather wide variation was seen in the examinees' reading proficiency, considered attributable in part to the relative degree of difference from English of the examinees' native language writing systems.

The scoring system for this section had originally included, in addition to a general "fluency" rating, separate ratings of "intonation" and "pronunciation"--a distinction which the project staff rating the tapes found very difficult to make. It was therefore decided to combine these two factors into a single scoring scale based on the overall degree of "intelligibility" and including a consideration of both intonation (pitch and stress) and pronunciation (realization of individual speech sounds).

Sentence Repetition (SENT REP)

As originally anticipated, this item type appeared to be very effective in discriminating between generally proficient and non-proficient speakers.

The most highly qualified speakers on the test as a whole were generally able to repeat most of the stimulus sentences quite accurately. The less proficient examinees, on the other hand, were observed to "lose track" of the sentences at a much earlier point and as a result would often not even attempt many of the later sentences in the series.

The initially specified scoring procedure for this section, which required the rater to judge whether "less than half" or "more than half" of the entire sentence had been reproduced accurately, was found difficult to apply consistently because of variation among scorers in their perception of what properly constituted "half" of the sentence. For this reason, the scoring system was modified by dividing each sentence into two separately evaluated segments. One to three points were awarded, depending on whether none, one, or both of the sentence segments were repeated exactly by the examinee. No response or isolated words were scored as zero.

An additional change based on the clinical results was to increase the number of questions of this type from 10 to 15, and to eliminate from both forms of the test some of the easiest items (e.g., "He's looking for his notebook"), which virtually every examinee was able to repeat correctly and which therefore provided little useful score variation among examinees.

Combining Sentence Elements (SENT ELEM)

This item type appeared to operate very smoothly from the test administration standpoint, with the exception of a few individual questions in which previously unanticipated homographic influences were discovered. For example, in "She/forget/wind/watch," some examinees interpreted "wind" as the cloud-moving agent, rather than as the intended manual action and were thus at a loss to respond appropriately even though they might indeed have been able to do so for an analogous set of sentence elements which did not contain an ambiguity of this type. The few questions containing this type of flaw were deleted from the large-scale administration version of the test.

The initial scoring key was also modified to provide incremental credit on the basis of the number of correct expansions (i.e., insertion of appropriate connectives at the slash [/] points), rather than on whole-sentence "communicativeness" and "grammatical accuracy," which could not be as readily assessed in this rather highly controlled format as in the later, more open-ended sections of the test.

Grammatical Patterns (GRAM PAT)

As previously anticipated, this item type was found to permit both very efficient and highly objective scoring, based on a simple dichotomous judgment of whether or not the correct grammatical pattern had indeed been produced by the examinee. Although previously cited negative aspects of this item type were fully appreciated by the project staff, it was decided

that for purposes of additional data gathering and analysis, the "grammatical patterns" item type would be included in the large-scale administration of the test.

Description of Picture Sequence (PICT SEQ); Multiple Questions on Single Picture (MULT Q)

These two general-proficiency-oriented formats are quite closely interrelated in that both are based on pictorial material which the examinee is asked to describe; and both procedures (especially the picture sequence) offer the examinee an appreciably greater degree of freedom in responding than is true of the item types appearing earlier in the test. It was therefore considered quite important to maintain at least one of these formats in the large-scale administration form of the test.

With respect to ease and straightforwardness of test administration, the PICT SEQ and MULT Q formats appeared to be generally on a par: both were quite easily dealt with by the examinees in the clinical study, albeit with widely differing degrees of linguistic sophistication. Of the two approaches, the picture sequence format produced somewhat more extensive and more complex speech on the part of the examinees, since in the single picture format, the examinee is repetitively required to stop and listen for the next question in order to continue with the description of the picture. The single picture format, on the other hand, was considered to be potentially more reliably scored, since its scoring key provided for separate rating of each of the question/answer response sets, rather than a single rating based on the totality of the examinee's narration.

In order to obtain more extensive comparison data for these two item types, it was decided to include both in the large-scale test administration and ultimately to select one of the two for operational use on the basis of the more detailed and more extensive information to be provided.

Simulated Telephone Conversation (TEL CONV)

Because of the rather complicated instructions required to familiarize the examinee with the testing procedures in this section and to adequately characterize the responses he or she is expected to make to the recorded voice, the simulated telephone conversation was found to raise a number of administration difficulties. One problem was a tendency for some of the examinees to turn back and forth in the test booklet between the paragraph description of the conversation and the printed telegraphic cues in an attempt to glean as much response information as possible from both sources, rather than to rely solely on the telegraphic cues as had originally been intended.

A second fairly serious problem was for the examinee to deviate from the intended sequence of responses, either by combining two separate items of information into a single reply or by failing to give a response before

the next statement/question in the sequence was spoken on the tape. In both instances, the naturalness and logical progression of the "conversation" would be at least temporarily lost--a situation which not only proved somewhat disconcerting to the examinee but also led to confusion on the part of the raters, who were being asked to evaluate each intended response as a separate entity.

Because of continuing interest in the telephone conversation format from the viewpoint of its relevance to real-life communication situations, this section of the test was retained for the large-scale administration, but with certain simplifications in the test directions and changes in the phrasing of the telegraphic cues which were intended to alleviate the examinee response difficulties encountered in the clinical study.

Persuasive Speech (PERSUAS)

This last section of the test was found to demonstrate essentially the same administration problems as the simulated telephone conversation, including the complexity (and attendant potential for confusion) of the test directions and the difficulty of maintaining the proper sequence of responding on the part of the examinees.

A further difficulty involved differences in the examinees' own reactions to the sociocultural situations represented in the "persuasive" format. One of the two situations especially--in which the examinee plays the role of a student who complains about and ultimately demands a refund for an unsatisfactory cafeteria meal--provoked widely differing, and most probably culturally-related, psychological reactions on the part of the examinees. For example, one of the Japanese examinees stated in post-test discussions that she would never "make such a scene" in a public setting and that as a result the testing task had been quite unnatural and somewhat disconcerting to her.

Because of the difficulties associated with the "persuasive speech" procedure, and in light of the fact that the somewhat similar "telephone conversation" format was to be included in the large-scale administration, it was decided by the project staff that the "persuasive" item type would not be dealt with beyond the initial clinical testing phase.

Sentence Completion (SENT COMPL)

Based on staff discussions following the clinical administration, a further item format not initially tested in the clinical study was proposed and developed for the large-scale administration. This "sentence completion" (SENT COMPL) format consisted of a series of printed partial sentences. For each, the examinee was to study the partial sentence silently and then respond by saying a full sentence incorporating both the beginning segment and his or her own completion:

Example:

(Printed in Test Booklet)

If it rains tomorrow...

(Examinee Response)

If it rains tomorrow, I'll stay
home/I will read a book/I will
not go the beach (etc.).

The SENT COMPL format was intended to provide a reasonable degree of latitude for the examinee's response and at the same time to permit fairly uniform and consistent scoring. Another positive aspect of this format was the possibility of testing across a wide range of proficiency levels by suitably varying the first element of the to-be-completed sentence. For example, whereas "If it rains tomorrow..." could be easily completed by examinees at a rather low proficiency level, an acceptable completion of a more linguistically complex sentence element such as "It has often been the case..." would require a much more sophisticated command of the language.

Content Outline of Revised Test Forms

The final inventory of question types included in the revised speaking test forms for the large-scale administration, together with the running times for each section, number of stimulus questions, and related information, is shown in Table 5. As with the initial clinical tests, two separate forms of the large-scale test were prepared, each containing appreciably more material than would ultimately be used in an operational test form. This approach allowed for considerable flexibility in final test specification, including both the selection of item types and, within item types, individual items (across both forms) showing the most suitable linguistic and statistical characteristics.

LARGE-SCALE ADMINISTRATION OF REVISED TEST FORMS

The final large-scale test administration following the initial clinical study was carried out at the University of Florida, Gainesville, with a few additional examinees subsequently added from the University of Pennsylvania in Philadelphia.

At both the Gainesville and Philadelphia sites, the tests and other measures administered to each examinee consisted of the following:

- (1) a recent form of the regular (three-section) TOEFL,
- (2) a background information/self-appraisal of speaking ability questionnaire,
- (3) an FSI-type speaking proficiency interview,

Table 5

Item Type Content of Revised Test Forms

<u>Section</u>	<u>Content of Section</u>	<u>Directions</u>	<u>Running Time*</u>	<u>Total</u>
			<u>Items</u>	
AUTOBIOG	7 questions	19 sec.	104 sec.	123 sec.
SV-N	32 pictured nouns	54	196	250
RD ALOUD	single reading passage	98	66	164
SENT REP	15 questions for repetition	52	199	251
SENT ELEM	12 elements for expansion	98	152	250
GRAM PAT	3 patterns; 5 items per pattern	246	165	411
SENT COMPL	10 sentences for completion	66	118	184
PICT SEQ	panel of 6 pictures for continuous narrative	70	62	132
MULT Q	6 questions on single picture	44	102	146
TEL CONV	telephone conversation with 7 anticipated replies by examinee	137	110	247

*Times shown are for Form 1 of the test; Form 2 timing is essentially identical.

- (4) two separate experimental speaking tests incorporating the testing formats and item types retained for further study following the earlier clinical administration.

For the Gainesville testing, arrangements were made with the director of the English Language Institute (ELI) to obtain the participation of approximately 150 non-native speakers of English who were taking courses at the Institute either simultaneously with or prior to enrolling in other subject matter courses at the university. Efforts were made to assemble a group of examinees having a very wide range of speaking ability and representing a large number of different mother tongue backgrounds.

Examinees were not paid for their participation in the study but were provided an official (institutional) TOEFL score without charge as well as the opportunity to obtain the results of their FSI interview. These were found to be quite effective incentives and it proved easily possible for the ELI staff to obtain the desired number of participants.

The main testing sessions in Gainesville were conducted over a three-day period in late October, 1978. Approximately one week beforehand, the ELI staff administered the TOEFL to all participants under standard administration conditions. During the first two days of the main testing, FSI-type interviews were administered individually to each participant. The interviews were conducted by three ETS staff members highly qualified in the interviewing process, together with three ELI staff who had previously undergone a one-day intensive training program in interviewing techniques conducted by the senior author. This training included discussion of the nature and operation of the FSI interview, observation of example interviews, and practice interviewing by each participant.

For the FSI interviewing, examinees arrived in groups of six at 30-minute intervals and were randomly assigned to one of the six interviewers. All interviews were conducted in quiet classrooms, with only the interviewer and interviewee present in each case. Each interview was tape recorded using the same recording arrangement (cassette recorder and two lapel microphones) employed for the earlier scoring reliability study. Interviews were of the "full length" type (i.e., not intentionally abbreviated as for the "short" interviews in the scoring reliability study). All interviews were completed by the afternoon of the second day of testing.

The experimental speaking test materials were administered during the third day. The examinee group was arbitrarily divided into two sections of approximately 70 students each--the practical capacity of the language laboratory facilities at the testing site.

Immediately prior to taking the first form of the speaking test the examinees were asked to complete a short background and self-appraisal questionnaire. As suggested by the results of the earlier scoring reliability study, the speaking self-appraisal section of the questionnaire was revised by considerably simplifying the language used in the six verbal competence descriptions and by providing an opportunity for the examinees

to mark the statement "my ability is in between these two descriptions" when they considered this a more accurate appraisal than selecting either the higher or lower category.

This "in-between" statement did not precisely match the official definition of "plus" values, which requires that the proficiency demonstrated "substantially exceed" the competence expected at the lower level in order to merit this attribution. It was considered an unfeasibly complex operation to attempt to instruct the examinees in the official meaning and use of "plus" ratings, and it was therefore decided to adopt the "in-between" description which, although less precise than the "plus" definition, was expected to allow for a greater degree of latitude and sensitivity in self-rating than would be provided by the simple level categories themselves. The text of the revised questionnaire is shown in Appendix B.

For each of the two examinee groups, completion of the questionnaire was followed immediately by administration of Form 1 or Form 2 of the experimental speaking test (depending on group membership) and, following a short rest break, the other form of the test. At each language laboratory booth position, examinees listened through padded earphones to the master test tape, which was distributed to the booths from the main laboratory console. Examinee responses were recorded on individual cassettes at each booth position by means of microphones integrated with the student headsets.

As with the clinical testing, the examinees did not as a general matter appear to be especially bothered by the overall length of the testing or negatively affected in their approach to the testing task. Indeed, in informal conversations with project staff, a number of examinees indicated considerable interest in the testing process and their own performance on the various sections of the test.

The obtained distribution of both TOEFL scores and FSI interview scores for the Gainesville examinees was found to be somewhat positively skewed. In order to obtain an additional number of examinees at the higher end of the proficiency range, a second test administration was held in early December, 1978 at the University of Pennsylvania. A total of 22 undergraduate or graduate level ESL students participated, previously selected by the university ESL staff to represent the most highly qualified students within the English program.

The Philadelphia testing session followed essentially the same procedures and scheduling as the Gainesville administration, and involved administration of a current TOEFL form about one week prior to the main testing, on-site FSI interviewing conducted by ETS project staff, administration of the same background/self-appraisal questionnaire used for the Gainesville testing, and language laboratory administration of the two experimental speaking test forms. Because of the relatively small number of examinees, only a single laboratory group was used for the experimental speaking tests, with Form 1 being administered first, followed after a short rest break by Form 2.

Scoring Procedures

Scoring of the completed FSI interviews was carried out by a total of three raters previously trained in this procedure by ETS staff in connection with another interview-based testing program in which the raters had been participating for about two and a half years at the time of the Gainesville testing.

In order to provide additional scoring detail beyond the overall "level" score itself (i.e., the single rating for each examinee of 2, 2+, 3, etc.), the FSI raters were also asked to evaluate the examinee's performance in each of the five language areas of: pronunciation, grammar, vocabulary, fluency, and comprehension. This rating was made on a two-dimensional grid by circling, for each of the five areas, one of five verbal descriptions of proficiency in that language area.

For example, the grid descriptions provided for the "vocabulary" area, in order of increasing competence, were:

Adequate only for survival, travel, and basic courtesy needs.

Adequate for simple social conversation and routine job needs.

Adequate for participation in all general conversation and for professional discussions in a special field.

Professional and general vocabulary broad and precise, appropriate to occasion.

Equal to vocabulary of an educated native speaker.

For each of the five areas, the raters were instructed to mark "on the line" between two verbal descriptions if the examinee's performance was considered to fall between these two descriptions. This procedure provided for a ten-point rating on each of the five language areas.

As a third and final step in the interview rating procedure for each examinee, the rater was asked to mark a separate list of "descriptive statements" for each of the five language areas. These statements were not part of the regular FSI rating system but were developed separately by project staff with the goal of expressing, in somewhat more operational terms, the same general kinds of discriminations being sought in the FSI grid. As an example, the descriptive statements for the area of "vocabulary" are shown below:

Vocabulary inadequate for even the simplest conversation.

Vocabulary restricted to stock phrases and basic survival terms.

Limitations of vocabulary hinder discussion of some general topics.

Vocabulary permits continuous conversation on a variety of topics, with some circumlocutions.

Vocabulary adequate to discuss any proposed topic, but occasional slight imprecision in choice of words or use of idioms.

For any topic broached, vocabulary as extensive and precise as that of an educated native speaker who is not a specialist in the area being discussed.

For this third rating technique, the number of descriptive statements (and hence, possible rating choices) varied by area as follows: pronunciation, 5; grammar, 6; vocabulary, 6; fluency, 4; comprehension, 6; there was no provision for "on the line" ratings.

Appendixes C, D, and E show the verbal criteria for the official FSI levels, the "grid" ratings, and the "descriptive statements" ratings, respectively.

In preparation for the scoring of the experimental speaking test tapes, a two-day discussion and training session was held to review the scoring system in detail and practice its application to a number of sample tapes. This session was led by the two report authors and attended by seven prospective scorers, including three ETS staff members and four outside individuals. All seven participants were highly familiar with TOEFL program operations and most had served as item writers for TOEFL as well as in other related testing activities at ETS.

During the first day of training, the purpose and rationale of the speaking test study was described in detail and a brief history of the study given. This was followed by a review of the scoring key developed for the clinical testing and the suggested modifications of the key based on the clinical testing results.

Several speaking test tapes from the Gainesville administration were then listened to and scored on a group basis, with further discussion of the revised scoring key. Further slight changes in the key were made on the basis of these discussions. Each participant was then given a "homework" assignment of additional tapes to be rated. These were discussed during the second day's training session and the scoring key, with one or two minor rephrasings of particular scoring criteria, was adopted as the formal, uniform scale for scoring the Gainesville (and subsequently, Philadelphia) speaking test tapes. The final version of the scoring key is shown in Appendix F.

DATA ANALYSIS AND RESULTS

The basic strategy for analyzing the large-scale administration data was to relate examinee performance on each experimental test item to performance on both the regular TOEFL and the FSI criterion interview.

Once a covariance matrix had been developed interrelating the above variables, each of the following analyses was carried out:

- (1) degree of whole-test (sum of items) correlation between each of the two experimental speaking tests--in the form in which they were administered in Gainesville/Philadelphia--and the FSI criterion interview (as well as TOEFL scores);
- (2) extent of correlation between item totals of the various item-type sections of the experimental tests (e.g., AUTOBIOG, RD ALOUD) and whole-test or subsection scores on FSI and TOEFL;
- (3) extent of correlation between each individual item in the experimental tests and whole-test or subsection scores on FSI and TOEFL.

The first type of analysis provided a general indication of the correlational performance of a speaking test consisting of the totality of items administered in the study (i.e., before any deliberate selection of the "better" item types or individual items had been made). The second analysis provided comparative information on the performance of the different item-type sections within the test. This information (together with simultaneous consideration of the linguistic and administrative strengths or drawbacks of the item types involved) aided in making informed decisions on those particular sections to be included in an eventual operational test form.

The third analysis made it possible, within any given item-type section, to zero in on those particular items showing the most substantial correlations with the criterion speaking test (FSI) scores. These items would be given first attention in the selection of items for an operational test, provided, of course, that they also met acceptable standards from linguistic and administrative standpoints.

The procedures followed in developing the necessary intercorrelation (covariance) matrix and carrying out each of the three types of analysis are described more fully below.

Development of Covariance Matrix

After the raw scoring data for both the FSI interview and the experimental speaking tests were received from the raters, they were checked for completeness and keypunched for entry into the computer programs. A few instances of response cassette tape breakage and a language laboratory transmission failure during a portion of the Philadelphia testing resulted in a small amount of missing data. The complex reliability design for the tape scoring yielded multiple scores for each examinee, usually by different raters, but in some cases by the same rater on a different day, for one-half of the speaking test tapes of each form, and for one half of the FSI interview tapes. These second sets of item scorings, identified by examinee and scorer, were keypunched, together with the item scorings of both speaking test forms, questionnaire responses, and FSI scorings and

ratings for each examinee. Frequencies, means, and standard deviations were obtained for each item in the data file. Where available, multiple scores on the same item were combined, and the examinee was assigned the resulting average score for that item. This procedure was considered to yield scoring reliability figures slightly lower than that those which would be expected in an operational speaking test program, since in the latter case all speaking test tapes (not simply a 50-percent sample) would routinely be scored by two raters.

Data for one section of the speaking test (Spoken Vocabulary--Nouns) were deliberately omitted from the analysis because during the speaking test administration in Gainesville it was noticed that even in spite of the sound-absorbing booths and padded headsets used in the language laboratory, it appeared possible for the examinees to "cheat" on this section by simply listening for the response of adjacent examinees before answering. Indeed, a distinctly antiphonal response pattern was observed to develop, with a large number of examinees delaying their responses until the group who knew the word had spoken. (This was not a problem in the other test sections, since individual variations in rate of speech and in choice of words reduced the background responses to an unintelligible murmur.) Since there would be no way to guard against such an occurrence in an operational program, the item type was removed from further consideration and analysis.

Experimental speaking test data entered into the covariance matrix for each of the two forms consisted of 97 coded item responses (excluding the SV-N items). TOEFL data included examinee scores on the Listening Comprehension, Reading Comprehension, and Structure and Vocabulary subsections (designated as TOEFL I, II, and III) and TOEFL total score (TOEFL T). FSI data consisted of: (1) the official total FSI rating on the 0-5 scale (FSI T) (see Appendix C); (2) separate "grid"-based ratings for pronunciation (FSI PRON), grammatical accuracy (FSI GRAM), vocabulary (FSI VOC), fluency (FSI FLU), and comprehension (FSI COMP) (Appendix D); and (3) similar ratings for each of the same five language areas based on the "descriptive statements" and denoted as FSI P, FSI G, FSI V, FSI F, and FSI C (Appendix E). In addition, selected combined descriptors for these five areas were calculated and entered into the covariance program, resulting in a final matrix of 211 x 211 elements.

Of 165 examinees with TOEFL scores and 164 with FSI interviews, sufficiently complete experimental speaking test response records for use in the analysis were available for 155 examinees. For each item on Form 1 of the experimental speaking test, records for 145-147 examinees were available. For each item on Form 2, complete data for 147-155 examinees were available through item 61 (the last item of the SENT COMPL item type). Beginning with item 62 on Form 2 (the first item of the PICT SEQ type), complete responses dropped to the range of 132-139, due to a transmission problem in the language laboratory during the Philadelphia testing. These missing data were replaced by means derived from the available data for each item. Means and standard deviations for total FSI scores, TOEFL scores, and background questionnaire variables are shown in Table 6.

Table 6

Means and Standard Deviations for
Criterion and Background Variables
(155 subjects)

	<u>Range</u>	<u>Mean</u>	<u>S.D.</u>
FSI	0-5	2.23	.73
Self-rating	0-5	2.40	.80
TOEFL I	0-80	45.33	8.32
TOEFL II+III	0-160	85.05	15.65
TOEFL T	200-800	434.44	73.92
Months in U.S.		4.42	6.28
Years English Study		5.82	3.41

Table 7 shows the correlations among the TOEFL and FSI variables and the total scores for each of the two forms of the experimental speaking test (ST1, ST2). Inspection of this matrix reveals that the TOEFL Listening Comprehension score (TOEFL I) correlates less strongly with the other two subsections of TOEFL than the level at which they relate to each other, but does correlate as strongly with the experimental speaking test scores as do the other two. TOEFL Total shows the highest correlations (.59 and .58) with the speaking test ratings. The total FSI rating exhibits a correlation of .53 with the Form 1 speaking test score, and .46 with the Form 2 score. This second correlation is in all probability depressed because of a moderate restriction in range resulting from the previously-described loss of data for the latter part of the Form 2 test for the higher-ability Philadelphia group. This pattern of lower FSI correlations for ST2 than for ST1 is maintained for most of the FSI "grid" and "descriptive statements" ratings as well.

It may be noted that the underlined correlations in Table 7 provide a multimethod reliability estimate of the FSI "grid" and "descriptive statement" ratings. Their values suggest that grammar can be rated quite reliably ($\underline{r} = .86$), as can vocabulary, fluency, and comprehension ($\underline{r} = .82, .81, .79$), but that there is considerably less scoring consistency in rating pronunciation ($\underline{r} = .67$).

The Table 7 data also indicate that the experimental speaking test total ratings did not correlate markedly higher with the TOEFL Listening Comprehension subsection than with the other two, more highly reading-based subsections (Reading Comprehension; Structure and Vocabulary). By the same token, the speaking test total ratings did not exhibit higher correlation with FSI grid-measured fluency, pronunciation, or comprehension ratings than with the grid-measured grammar or vocabulary ratings. Since the latter two skill areas are already measured with considerable adequacy by the current TOEFL, it was considered desirable for speaking test development purposes to concentrate on the identification of those item type sections or individual items that were found to be heavily weighted on fluency, pronunciation, and comprehension, as measured by the corresponding

Table 7

Intercorrelations of TOEFL, FSI, and Experimental Speaking Test Scores

	TOEFL I	TOEFL II	TOEFL III	TOEFL T	FSI T	FSI PRON	FSI GRAM	FSI VOC	FSI FLU
TOEFL I	1.00	.71	.74	.90	.65	.38	.62	.62	.53
TOEFL II	.71	1.00	.77	.91	.58	.34	.59	.55	.49
TOEFL III	.74	.77	1.00	.92	.61	.33	.55	.60	.53
TOEFL T	.90	.91	.92	1.00	.68	.38	.65	.65	.57
FSI T	.65	.58	.61	.68	1.00	.62	.85	.92	.81
FSI PRON	.38	.34	.33	.38	.62	1.00	.69	.57	.62
FSI GRAM	.62	.59	.55	.65	.85	.69	1.00	.76	.67
FSI VOC	.62	.55	.60	.65	.92	.57	.76	1.00	.81
FSI FLU	.53	.49	.54	.57	.81	.62	.67	.80	1.00
FSI COMP	.61	.51	.53	.61	.87	.60	.70	.83	.82
FSI P	.44	.45	.40	.47	.70	.67	.56	.64	.66
FSI G	.60	.58	.55	.63	.85	.66	.86	.77	.67
FSI V	.54	.43	.48	.53	.81	.55	.67	.82	.80
FSI F	.52	.44	.50	.54	.74	.52	.58	.76	.81
FSI C	.60	.52	.55	.61	.80	.48	.64	.76	.73

	FSI COMP	FSI P	FSI G	FSI V	FSI F	FSI C	ST 1	ST 2
TOEFL I	.61	.44	.60	.54	.52	.60	.52	.55
TOEFL II	.51	.45	.58	.43	.44	.52	.54	.50
TOEFL III	.53	.40	.54	.48	.50	.55	.53	.53
TOEFL T	.61	.47	.63	.53	.54	.61	.59	.58
FSI T	.87	.70	.65	.81	.74	.80	.53	.46
FSI PRON	.60	.67	.66	.55	.52	.48	.32	.22
FSI GRAM	.70	.56	.86	.67	.58	.64	.50	.40
FSI VOC	.83	.64	.77	.82	.76	.76	.51	.48
FSI FLU	.82	.66	.67	.80	.81	.73	.45	.47
FSI COMP	1.00	.69	.77	.82	.73	.79	.50	.46
FSI P	.69	1.00	.61	.68	.62	.67	.41	.31
FSI G	.77	.61	1.00	.67	.60	.67	.51	.44
FSI V	.82	.68	.67	1.00	.78	.78	.48	.44
FSI F	.73	.62	.60	.78	1.00	.75	.48	.44
FSI C	.79	.67	.67	.78	.75	1.00	.49	.49
ST1							1.00	.87

FSI subscores, but that had at the same time relatively less strong correlations with TOEFL scores, especially TOEFL II and III scores.

This selection strategy implied that certain item types or individual items from the experimental speaking test forms that were found to have substantial correlations with FSI total or subscores would have to be discarded if they simultaneously exhibited very strong relationships to TOEFL scores. However, this rather stringent approach to final item selection was considered desirable if the final speaking test was to tap important aspects of spoken communication while at the same time overlapping as minimally as possible with language skills already measured by the current TOEFL format.

Item-Type Analyses

Detailed item type (as well as individual item) analyses of the speaking test data were greatly facilitated through the use of the SCALAR item analysis program (Rogers and Swinton, 1977), developed to permit the "real-time" analysis and modification of questionnaires and tests containing more than one scale and employing items with shared method variance not necessarily relevant to the intended constructs to be measured. The SCALAR program makes it possible to add and delete items and scales in an interactive mode, and to observe immediately the impact of such changes on item-scale correlations (with the scale in which the item is included, with other subscales, and with total score) on mean difficulty, standard deviation, reliability, interscale correlations, and on correlations of scales with external criteria.

In the present study, the use of SCALAR made it possible to examine, for example, the value of combining grammar ratings across several items separately from fluency ratings for the same items--combinations which would not have resulted from exploratory analytic procedures, such as factor analysis, because of the strong interdependencies of ratings based on the same passage. This ability to maintain a priori structural specifications in performing SCALAR analyses carries with it the price of suboptimizing rather than optimizing reliability; and because of the trial-and-error nature of program use, the temptation to try "just one more change" in item selection may tend to capitalize on error variance after the practical limit of correlational improvement has been reached, a temptation which must be steadfastly resisted by the investigator.

The first item selection pass using the SCALAR program identified those 48 items from each form having the highest correlations with the total FSI score. The correlations of the totals of these selected items with the FSI criterion--.71 for the Form 1 items, and .68 for the Form 2 items--were acceptable, but the correlations of these same item totals with TOEFL Total--.81 and .79--were even higher, approaching the intercorrelation of the two experimental test scales: .87. This initial finding was not particularly surprising because the test-retest reliability of

TOEFL scores is in all probability higher than the inter-interview (test-retest) reliability of the FSI, and, as a consequence, the observed result might be expected even if the true score correlations of the speaking test with FSI were as high as or even higher than the true score correlations of the speaking test with TOEFL. However, it was desired to do better than that if possible, in the sense of achieving an even stronger observed correlation with FSI than with TOEFL. To this end, the 48 selected speaking test items for each form were combined into a single 96-item scale for further analysis. Item type correlations with FSI were for this second analysis contrasted with correlations with a derived "TOEFL II+III" score rather than the entire TOEFL total, it being reasoned that high correlations with TOEFL I were not undesirable in a measure of speaking competence. In addition, wherever there were multiple scoring criteria for a given item type (for example, the "pronunciation," "grammar," and "fluency" ratings for the PICT SEQ section), each of these was analyzed separately so that the relative correlational performance of the different scoring criteria could be determined. Table 8 gives the median item-criterion correlations of the experimental item types from Forms 1 and 2 with the FSI Total and the TOEFL II+III criteria.

As shown in Table 8, of the 17 pairs of median correlations, TEL CONV (as scored for "communication") correlated higher with FSI than with TOEFL II+III for both speaking test forms, and six other item types--AUTOBIOG, PICT SEQ (PRON), PICT SEQ (FLU), MULT Q (FLU), TEL CONV (GRAM), and TEL CONV (FLU)--showed greater median correlations with FSI than with TOEFL II+III for one of the two forms. For GRAM PAT (A) and (B), the correlations with FSI and TOEFL II+III were identical for one form of the test, with higher correlations with TOEFL II+III for the other test form. For the remaining eight item types--RD ALOUD (PRON), RD ALOUD (FLU), SENT REP, SENT ELEM, GRAM PAT (C), SENT COMPL, PICT SEQ (GRAM), and MULT Q (GRAM)--both Form 1 and Form 2 median correlations with TOEFL II+III exceeded the correlations with FSI. The item type selection strategy thus became that of minimizing the excess of TOEFL correlations over correlations with FSI, consistent with retaining items which were good predictors of the FSI score.

Based on the Table 8 results, together with simultaneous consideration of the linguistic content and administrative characteristics of each item type, decisions were made as to the particular item types to be recommended for inclusion in a prototype "Test of Spoken English" (TSE). Once these item types had been chosen, a further analysis of individual items within each selected item type was conducted to select those particular items that would make up the prototype test. Decisions as to item types were as follows:

Table 8

Median Correlations of Experimental Speaking Test Items
with FSI and with TOEFL II+III

<u>Item Type</u>	<u>Form</u>	<u>N</u>	<u>rFSI</u>	<u>rTOEFL II+III</u>
AUTOBIOG	1	7	.27	.17
	2	7	.21	.21
RD ALOUD (PRON)*	1	1	.51	.59
	2	1	.48	.50
RD ALOUD (FLU)	1	1	.39	.55
	2	1	.42	.55
SENT REP	1	15	.41	.52
	2	15	.40	.44
SENT ELEM	1	12	.41	.48
	2	12	.39	.48
GRAM PAT (A)**	1	5	.35	.48
	2	5	.21	.21
GRAM PAT (B)	1	5	.41	.41
	2	5	.21	.28
GRAM PAT (C)	1	5	.30	.43
	2	5	.18	.46
SENT COMPL	1	10	.38	.43
	2	10	.37	.46
PICT SEQ (PRON)	1	1	.55	.49
	2	1	.55	.57
PICT SEQ (GRAM)	1	1	.55	.58
	2	1	.45	.50
PICT SEQ (FLU)	1	1	.54	.47
	2	1	.50	.50
MULT Q (GRAM)	1	6	.33	.35
	2	6	.35	.40
MULT Q (FLU)	1	6	.39	.35
	2	6	.42	.45
TEL CONV (COM)	1	7	.29	.25
	2	7	.27	.24
TEL CONV (GRAM)	1	7	.39	.28
	2	7	.31	.38
TEL CONV (FLU)	1	7	.44	.36
	2	7	.38	.39

*Refers to responses in this section as scored for Pronunciation, Grammar, etc.

**Letters refer to each of three separate patterns tested in this section.

Autobiographical Questions (AUTOBIOG)

This item type showed very low median correlations with both the FSI and TOEFL II+III criteria. This is in all probability due to the largely "memorized" nature of the information requested in this section, which even less proficient examinees would be able to handle reasonably well and which would therefore result in a relatively restricted score range. Although the autobiographical questions would be expected to provide little useful score variation in an operational test, they used testing time very efficiently and were also considered quite non-threatening to the examinee since they dealt with familiar topics in a quasi-conversational setting. In view of these positive aspects, it was decided to retain the autobiographical questions as a non-scored warm-up exercise at the beginning of the test.

Reading Aloud (RD ALOUD)

The "pronunciation" rating for the reading aloud section showed substantial correlations (.51, .48) with the FSI criterion across both test forms, but even higher correlations (.59, .50) with TOEFL II+III. The "fluency" rating for reading aloud showed almost equally strong correlations with TOEFL II+III, but slightly weaker correlations with FSI. Other available "fluency" ratings (from the PICT SEQ, MULT Q, and TEL CONV sections) showed generally higher correlations with FSI than with TOEFL II+III. Fluency in reading aloud is clearly more strongly influenced by receptive grammatical and lexical competence than by the unprompted speaking skills involved in the other, more free-response, fluency measures, and as such could be expected to relate more strongly to the skills measured by the regular TOEFL than to the "free-production" fluency at issue in the FSI interview.

In view of both the real-life validity of reading aloud as a component of academically-oriented speaking proficiency and the high absolute correlations of the RD ALOUD section with the FSI scores, it was decided to retain this section (scored for both pronunciation and fluency) in the prototype TSE, despite its substantial correlational relationship to TOEFL II+III.

Sentence Repetition (SENT REP)

This item type, initially a strong candidate for inclusion on psycholinguistic grounds, showed fairly high median correlations (.41, .40 for Forms 1 and 2) with FSI, but even stronger (.52, .44) correlations with TOEFL II+III. As previously discussed, the sentence repetition format was considered to possess relatively low face validity and, in view of both face-validity and relative FSI/TOEFL correlations, was reluctantly omitted from further consideration, despite its theoretical appeal as a testing technique.

Combining Sentence Elements (SENT ELEM)

This item type was initially considered to have relatively low face validity (in that slash-separated sentence parts would not ordinarily be encountered in real-life language situations), but it had been included in both the clinical and large-scale administration tests as a "bridge" between the highly-controlled reading aloud and sentence repetition sections and the freer-response sections appearing later in the test. Correlational levels for this section with the FSI were found to be relatively high (.41, .39) but, unfortunately, were substantially higher with TOEFL II+III (.48 for both test forms). In light of the statistical results, and in comparing the SENT ELEM format to other item type sections showing higher face validity and/or more favorable correlational relationships with FSI and TOEFL II+III, it was decided that this item type did not warrant inclusion in the prototype TSE.

Grammatical Patterns (GRAM PAT)

Across all six of the grammatical patterns tested in the two forms of the speaking test, the median correlations of these items with the FSI interview were identical to the corresponding correlations with TOEFL II+III in two instances and appreciably lower in the four remaining instances: for example, .48 to .35 for GRAM PAT (A) (Form 1), and .46 to .18 for GRAM PAT (C) (Form 2). The poor overall predictive ability for FSI and, at the same time, relatively high correlations with TOEFL II+III provided no statistical justification for including the GRAM PAT format in the prototype test; and to this were added the previously-identified problems of adequate pattern sampling within a test of reasonable length. This item type was removed from further consideration on both statistical and test-content grounds.

Sentence Completion (SENT COMPL)

This relatively open-ended item type showed slightly lower correlations with FSI and relatively very slightly lower correlations with TOEFL II+III than did the SENT ELEM item type. However, the easier (i.e., less grammatically complex) items of the SENT COMPL type were seen to have approximately equal correlations with FSI and with TOEFL II+III. In keeping with the goal of selecting, for the speaking test, items weighted generally on fluency and delegating the measurement of grammatical knowledge to the paper-and-pencil TOEFL, it was decided to retain the less grammatically demanding items of the SENT COMPL type for the prototype speaking test and to drop the more complex items showing relatively higher correlations with TOEFL II+III.

Picture Sequence (PICT SEQ)

This proved to be one of the most effective item types in the experimental test. Although the grammar ratings for PICT SEQ correlated (not

unexpectedly) somewhat higher with TOEFL II+III than with FSI, both the pronunciation and fluency ratings correlated more highly, on the average, with FSI. This was particularly true for Form 1, in which a wider range of speaking test scores was available. Not only were FSI vs. TOEFL II+III correlations favorable, but the absolute values of the median correlations with FSI (.45 to .55) were impressively high.

From a linguistic standpoint, the PICT SEQ format was considered highly open-ended and productive of fairly extensive responses on the part of qualified examinees. These factors, together with the encouraging statistical data, were considered to justify inclusion of this item type in the final prototype test.

Multiple Questions on Single Picture (MULT Q)

As was true of the grammar ratings for PICT SEQ, the grammar ratings for MULT Q correlated somewhat more highly with TOEFL II+III than with FSI. Correlations of the fluency ratings were about equal for TOEFL II+III and FSI and were sufficiently high to suggest a worthwhile contribution to total test validity. Although the MULT Q format was somewhat more restrictive than the PICT SEQ format (in that it requires the examinee to wait passively for each question rather than permitting a continuous and more spontaneous narrative), it was considered a worthwhile supplement to the PICT SEQ format and was also retained in the prototype TSE.

Simulated Telephone Conversation (TEL CONV)

The statistical performance of this item type showed relatively higher correlations with FSI than with TOEFL II+III for all three rating areas--communication, grammatical accuracy, and fluency--and on these grounds would have justified inclusion in the prototype test. However, the previously-cited difficulties in properly timing and integrating the recorded stimulus questions and statements with the anticipated examinee responses, together with the lengthy instructions required to properly orient the examinee to the testing task, were rather substantial negative factors. After considerable discussion, the project staff decided to delete the TEL CONV format from the final prototype test, primarily on grounds of administrative practicality rather than for any deficiency in statistical performance. This decision was facilitated somewhat by the recognition that two other relatively open-ended formats (PICT SEQ and MULT Q) were being retained in the prototype test and could be expected to provide for a useful degree of "free" expression on the part of the examinee, even in the absence of the simulated telephone conversation.

Individual Item Selection and Scaling

Once the general item types for inclusion in the prototype TSE had been determined, selection of specific items within each item type could proceed, again with the goal of maintaining the highest possible

correlations with the FSI rating while keeping speaking test/TOEFL correlations as low as possible and at least no higher than speaking test/FSI correlations. At this stage, correlations of the provisionally-selected speaking items with FSI fluency, pronunciation, comprehension, grammar and vocabulary ratings (as measured by the "grid" technique) were also examined, with a view to selecting items which tapped the first three of these rating areas as strongly as or more strongly than they did the more reliable grammar and vocabulary ratings.

After statistically analyzing several possible subsets of items drawn from the two speaking test forms, and also taking into account the need to provide a suitable content coverage and difficulty range for the test as a whole, a 20-element prototype Test of Spoken English (TSE), as outlined in Table 9, was selected. This prototype test showed the rather impressive internal consistency reliability of .91, and had equal correlations ($r = .68$) with FSI ratings and with TOEFL II+III. Its correlation with TOEFL I (Listening Comprehension) was also .68, and with TOEFL Total, .72. The prototype TSE showed equal correlations (.67) with the "grid"-based FSI pronunciation + fluency + comprehension (FSI P+F+C) rating, and with the FSI grammar + vocabulary (FSI G+V) rating.

From the above values, the multiple correlation predicting FSI from both TOEFL and the prototype speaking test is calculated as .731. Thus, the addition to the regular TOEFL of a prototype speaking test consisting of the elements shown in Table 9 would be expected to raise the percentage of predictable variance of FSI scores by about 8% (45.8% to 53.4%). By deleting all of the grammar ratings for PICT SEQ and MULT Q (with no change in the total number of stimulus items), correlations of the remaining 15-element total score with FSI, with TOEFL Listening, and with FSI P+F+C can be made to rise slightly, to .69, while correlations with TOEFL II+III remain at .68, and with FSI G+V ratings, at .67. However this miniscule improvement in discriminant validity comes at the price of a reduction in total speaking test reliability from .91 to .87; this change was not judged to be a net advantage.

Similarly, inspection of Table 9 reveals that several items, e.g., RD ALOUD (FLU) and SENT COMPL (2)(7), show higher correlations with TOEFL II+III than with FSI. Deleting these items would also improve discriminant validity slightly, but again at the cost of reliability.

For preliminary score-scaling analysis, it was decided to place the prototype TSE scores on a 0-50 scale. This scale would be anchored to the FSI scale at 2+ = 27, and to the extent that the shapes of the distributions were comparable across the two instruments, this procedure would yield scores in a metric easily translatable to an approximation of the familiar FSI scale.

Development of an equation for determining this scale was based on multiple regression, allowing those item types with strongest relationships to FSI (i.e., RD ALOUD (PRON) and PICT SEQ (PRON)) to be given more weight in the prediction equation. This equation was thus of the form: $FSI = 5.34 + 1.96 \text{ RD ALOUD} + .75 \text{ SENT COMPL}$, etc. Because of measurement

Table 9
Correlations of Prototype TSE Items
with Criterion Measures

Item*	TSE Total	FSI	TOEFL I	TOEFL II+III	TOEFL T	FSI P+F+C**	FSI G+V**
RD ALOUD(1)(PRON)	.55	.51	.59	.59	.62	.45	.49
RD ALOUD(1)(FLU)	.48	.39	.50	.55	.56	.35	.31
SENT COMPL(1)(4)	.50	.40	.33	.38	.38	.40	.38
SENT COMPL(1)(7)	.61	.47	.47	.53	.53	.43	.46
SENT COMPL(2)(1)	.49	.36	.34	.44	.43	.36	.41
SENT COMPL(2)(4)	.42	.33	.33	.29	.32	.39	.32
SENT COMPL(2)(7)	.37	.36	.41	.38	.41	.38	.32
SENT COMPL(2)(8)	.41	.31	.32	.27	.31	.33	.32
SENT COMPL(2)(9)	.42	.39	.34	.27	.31	.39	.40
PICT SEQ(1)(PRON)	.67	.55	.49	.49	.52	.53	.41
PICT SEQ(1)(GRAM)	.69	.55	.56	.58	.60	.44	.56
PICT SEQ(1)(FLU)	.74	.54	.52	.47	.52	.53	.54
MULT Q(1)(GRAM)(1)	.52	.34	.30	.33	.34	.33	.31
MULT Q(1)(GRAM)(2)	.57	.31	.33	.37	.37	.28	.34
MULT Q(1)(GRAM)(3)	.56	.35	.39	.37	.40	.30	.40
MULT Q(1)(GRAM)(6)	.65	.42	.41	.41	.43	.45	.43
MULT Q(1)(FLU)(1)	.55	.43	.37	.36	.38	.44	.44
MULT Q(1)(FLU)(2)	.60	.39	.37	.32	.35	.38	.41
MULT Q(1)(FLU)(3)	.59	.39	.41	.35	.39	.42	.37
MULT Q(1)(FLU)(6)	.67	.41	.45	.45	.48	.46	.43

*Legend: Following the item type designation (e.g., RD ALOUD), test form appears in parentheses (1;2). PRON, GRAM, and FLU refer to these particular section scorings. Where they appear, second numbers in parentheses denote item numbers within sections.

**As measured by FSI "grid" ratings.

Table 10

Regression Equation Predicting FSI Score from TSE Items

Constant = 5.3437

Multiple R = .715

Standard Error
of Estimate: 5.36

Item*	Range	Regression Weight	Standard Error of Weight	Contribution to R ²
RD ALOUD (1)(PRON)	0-4	1.9596	.8625	.0178
SENT COMPL (1)(4)	0-2	.7495	.7180	.0038
SENT COMPL (1)(7)	0-2	1.2475	1.0238	.0051
SENT COMPL (2)(1)	0-2	.6252	.7412	.0025
SENT COMPL (2)(4)	0-2	.8988	.7213	.0053
SENT COMPL (2)(7)	0-2	1.0141	.6228	.0091
SENT COMPL (2)(9)	0-2	1.0193	.7171	.0070
PICT SEQ (1)(PRON)	0-4	2.3843	1.1149	.0158
PICT SEQ (1)(FLU)	0-4	.7380	.8071	.0029
MULT Q (1)(FLU)(1)	0-3	1.3280	.6834	.0130
MULT Q (1)(GRAM)(2)	0-2	-1.4877	1.0832	.0065
MULT Q (1)(GRAM)(6)	0-2	1.1203	.8547	.0059

Regression Equation Predicting FSI Score from TSE Subscores

Constant = 4.7502

Multiple R = .702

Standard Error
of Estimate = 5.34

Item*	Range	Regression Weight	Standard Error of Weight	Contribution to R ²
RD ALOUD (1)(PRON)	0-4	1.5893	.8526	.0119
SENT COMPL (1+2)	0-14	.7864	.1909	.0581
PICT SEQ (1)(PRON)	0-4	1.7515	1.1804	.0075
PICT SEQ (1)(GRAM)	0-4	1.0228	1.1106	.0029
PICT SEQ (1)(FLU)	0-4	.5623	.8819	.0014
MULT Q (1)(FLU)	0-12	.3727	.2821	.0060

*See Table 9 for legend.

redundancy, not all items added significant additional predictive power after the best predicting items had been entered, and the initial equation made use of only half of the experimental items. As new items are developed to parallel the present experimental items and additional administration data are gathered, the prediction equation will be revised and less sample-dependent scoring weights will be derived. The initial regression equation developed in this study is given in Table 10: this equation was found to yield a multiple correlation with FSI total scores of .702.

The difficulty of achieving a TOEFL+TSE multiple correlation greater than .70-.72 with the FSI would appear to lie at least partly with the reliability of the FSI criterion, given the high parallel-form (Forms 1 and 2) correlations of the initial 48-element speaking test (.896), and the impressive internal consistency of the final 20-element prototype TSE (.906), as well as the high internal reliability of TOEFL I (.871) and TOEFL II+III (.895).

Although interrater reliability of the FSI seems satisfactory, it will not be possible to assess the degree to which TSE scores predict reliable FSI variance without additional well-controlled inter-interview (i.e., test-retest) reliability studies of the FSI technique over the range of speaking abilities actually encountered operationally. The value of .696 obtained from the Table 3 data for "long" vs. "short" interviews may be considered a lower bound for inter-interview reliability of full-length interviews, but the apparent ceiling on FSI predictability using both TOEFL and TSE predictors suggests that this lower bound may not be far below the actual value.

FURTHER RESEARCH AND DEVELOPMENT ACTIVITIES

The activities reported in the present study involved the carrying out of an experimentally-based, iterative process of development, trial administration, and preliminary selection of a number of item types and associated scoring procedures intended to provide a face-valid and reliable measure of English speaking proficiency within the operational framework of the TOEFL program. The final "product" of the study was a prototype Test of Spoken English submitted to the TOEFL program direction office for consideration for eventual operational use in the TOEFL program.

Subsequent to the completion of this study, approval has been obtained to carry out two additional speaking test-related studies. The first will involve the pilot testing, under actual operational conditions, of a slightly modified version of the TSE form developed in the initial study. The pilot administration study will determine the practicality of speaking test administration at representative TOEFL sites and will point out any procedural modifications required in test instructions, administration procedures, or other operational aspects of the test. At the same time, additional data on examinee performance will be gathered for further analysis and naming purposes.

Simultaneously with the pilot administration study, a second research project will be carried out, involving the comparison of scores obtained by non-native English speaking teaching assistants (and other graduate-level instructors) on the Test of Spoken English to ratings of their "classroom communication" ability as judged by native English-speaking students taking their courses. The intent of this study is to obtain use-referenced validity data for one important prospective application of TSE score results: the initial selection and/or appropriate instructional assignment of foreign teaching assistants with respect to empirically established levels of communicative ability in an academic setting.

Both of these studies are currently under way and on completion will be fully reported in the TOEFL research report series or through other publication means.

REFERENCES

- Carroll, J. B., Clark, J. L. D., Edwards, T. M., & Handrick, Fannie A. The foreign language attainments of language majors in the senior year: a survey conducted in U.S. colleges and universities. Cambridge, Mass.: Laboratory for Research in Instruction, Graduate School of Education, Harvard University, 1967.
- Clark, H. H., & Clark, E. Psychology and language: an introduction to psycholinguistics. New York: Harcourt Brace Jovanovich, 1977, 137-41.
- Clark, J. L. D. (Ed.). Direct testing of speaking proficiency: theory and application. Princeton, N.J.: Educational Testing Service, 1978.
- Clark, J. L. D. Direct vs. semi-direct tests of speaking ability. In E. J. Brière and F. B. Hinofotis (Eds.), Concepts in language testing: some recent studies. Washington, D. C.: Teachers of English to Speakers of Other Languages, 1979, 35-49.
- Clark, J. L. D. Foreign language testing: theory and practice. Philadelphia, Pa.: Center for Curriculum Development, 1972.
- Clark, J. L. D. Some considerations in preparing test questions. In Northeast Conference Reports of the Working Committees (W.C. Born, Ed.). Middlebury, Vt.: Northeast Conference on the Teaching of Foreign Languages, 1975, 99-121.
- Coste, D., Courtillon, J., Ferenczi, V., Martins-Baltar, M., Papo, E., & Roulet, E. Un niveau-seuil. Strasbourg: Council of Europe, 1976.
- Lado, R. Language testing: the construction and use of foreign language tests. London: Longmans, Green, 1961.
- Pike, L. W. An evaluation of alternative item formats for testing English as a foreign language. TOEFL Research Report No. 2. Princeton, N.J.: Educational Testing Service, June 1979.
- Rogers, A., & Swinton, S. SCALAR computer analysis program. Princeton, N.J.: Educational Testing Service, 1977.
- Sollenberger, H. E. Development and current use of the FSI oral interview test. In J. L. D. Clark (Ed.), Direct testing of speaking proficiency: theory and application. Princeton, N.J.: Educational Testing Service, 1978, 3-12.
- van Ek, J. The threshold level in a European unit/credit system for modern language learning by adults. Strasbourg: Council for Cultural Co-operation, Council of Europe, 1975.

Valette, R. M. Modern language testing: a handbook. New York: Harcourt Brace & World, 1967.

Wilds, C. P. The oral interview test. In R. L. Jones & B. Spolsky (Eds.), Testing language proficiency. Arlington, Va.: Center for Applied Linguistics, 1975, 29-44.

Appendix A

Initial Speaking Test Scoring Key

AUTOBIOGRAPHICAL

- 0 = no response or "single-word answer"
- 1 = communicates information adequately, but one or more grammatical flaws
- 2 = communicates information adequately, no grammatical error

SPOKEN VOCABULARY - NOUNS

- 0 = no response or incorrect identification of picture
- 1 = proper identification of picture, with phonemically accurate pronunciation

SPOKEN VOCABULARY - VERBS

- 0 = no response or incorrect identification of action
 - 1 = proper identification of action, with phonemically accurate pronunciation
- For each action involving a direct object:
- 0 = does not attempt or incorrectly handles modifier and direct object
 - 1 = correctly handles modifier and direct object

READING ALOUD

Score entire passage for:

Intonation

- 0 = no response or only isolated words
- 1 = serious intonation problems, very difficult to understand
- 2 = departures from native intonation occasionally obscure meaning
- 3 = slight departures from native intonation, do not obscure meaning
- 4 = intonation patterns native-like throughout

Pronunciation

- 0 = no response or only isolated words
- 1 = heavy non-native accent, numerous phonemic errors
- 2 = occasional phonemic errors, many phonetic errors
- 3 = occasional phonetic errors
- 4 = completely native pronunciation

Fluency

- 0 = no response or only isolated words
- 1 = numerous non-native pauses, fails to finish in time allotted
- 2 = several non-native pauses but finishes in time allotted
- 3 = native fluency except for occasional slight hesitancy
- 4 = native immediacy and ease of response

SENTENCE REPETITION

- 0 = no response
- 1 = less than half the sentence repeated accurately
- 2 = more than half the sentence repeated accurately but some errors
- 3 = exact repetition of entire sentence

COMBINING SENTENCE ELEMENTS

- 0 = no response or response does not adequately communicate message
- 1 = communicates message adequately with some grammatical errors
- 2 = communicates message adequately with completely correct grammar

GRAMMATICAL PATTERNS

- 0 = no response or response containing grammatical errors
- 1 = completely accurate response fulfilling requirements of pattern

DESCRIPTION OF PICTURE SEQUENCE

Separate score for:

Fluency

- 0 = story not attempted or virtually no response
- 1 = more extensive than 0, but hesitant and falling short of complete narration of critical aspects
- 2 = smooth but non-elaborated narration of critical aspects
- 3 = smooth and complete narration with some attempt at elaboration

Grammatical Accuracy

- 0 = no response or only isolated words
- 1 = frequent grammatical errors
- 2 = occasional grammatical errors
- 3 = very few errors, of a minor nature
- 4 = no grammatical errors

Pronunciation

- 0 = no response or only isolated words
- 1 = considerable phonemic errors
- 2 = some phonemic errors
- 3 = non-native accent but no phonemic errors
- 4 = indistinguishable from native speaker

Extent of Response

0-6 points (1 point for each picture description attempted)

MULTIPLE QUESTIONS ON SINGLE PICTURE

For each question, separate score for:

Fluency

- 0 = no response or only isolated words
- 1 = hesitant response, obvious groping for words
- 2 = slight non-native disfluency
- 3 = native immediacy and ease of response

Grammatical Accuracy

- 0 = no response or only isolated words
- 1 = grammatical errors
- 2 = no grammatical errors

SIMULATED TELEPHONE CONVERSATION

For each response:

- 0 = fails to convey message
 - 1 = conveys message
- Global rating of conversation:

Fluency

- 0 = no response or only isolated words
- 1 = hesitant response, obvious groping for words
- 2 = slight non-native disfluency
- 3 = native immediacy and ease of response

Grammatical Accuracy

- 0 = no response or only isolated words
- 1 = grammatical errors
- 2 = no grammatical errors

PERSUASIVE SPEECH

For each response:

- 0 = fails to convey message
- 1 = conveys message

Global rating of conversation:

Fluency

- 0 = no response or only isolated words
- 1 = hesitant response, obvious groping for words
- 2 = slight non-native disfluency
- 3 = native immediacy and ease of response

Grammatical Accuracy

- 0 = no response or only isolated words
- 1 = grammatical errors
- 2 = no grammatical errors

PLEASE PRINT CLEARLY

- [] I can ask and answer questions about very familiar subjects and can understand simple questions and statements if they are spoken slowly and sometimes repeated. My vocabulary is limited to basic needs (food, asking directions, greeting people, and so forth). I make many grammatical mistakes, but my teachers can usually understand me. I can order food in a restaurant, get a room in a hotel, ask directions on the street, and introduce myself to people.
- [] (My ability is in between these two descriptions.)
- [] I can talk with native speakers of English about myself and my family, my job or studies, hobbies, and current events. I can understand most conversations in English except when the speech is very fast. My grammar is fairly good but I make mistakes with complicated constructions. If I do not know the word for a particular thought or object, I can usually describe it by using other, easier words.
- [] (My ability is in between these two descriptions.)
- [] I can understand almost everything spoken by native English speakers. My vocabulary is good enough that I usually know most or all of the words for what I want to say. My grammar is good and any mistakes I make are usually with the more complicated constructions. I can pronounce English clearly but do not have a perfect accent.
- [] (My ability is in between these two descriptions.)
- [] I can understand native English speakers even when they are speaking quickly and using sophisticated or colloquial expressions. My vocabulary is very extensive, even for technical matters, and I can talk fluently and accurately about almost any subject that I know about. I make only a very few grammatical errors and my pronunciation is very good but not completely native.
- [] (My ability is in between these two descriptions.)
- [] My speech is exactly like that of an educated native speaker of English.

Please return the completed questionnaire to your supervisor. Thank you!

Appendix C
FSI Level Descriptions

Level

Verbal Descriptions

- 1 Able to satisfy routine travel needs and minimum courtesy requirements. Can ask and answer questions on topics very familiar to him; within the scope of his very limited language experience can understand simple questions and statements, allowing for slowed speech, repetition or paraphrase; speaking vocabulary inadequate to express anything but the most elementary needs; errors in pronunciation and grammar are frequent, but can be understood by a native speaker used to dealing with foreigners attempting to speak his language; while topics which are "very familiar" and elementary needs vary considerably from individual to individual, any person at the S-1 level should be able to order a simple meal, ask for shelter or lodging, ask and give simple directions, make purchases, and tell time.
- 2 Able to satisfy routine social demands and limited work requirements. Can handle with confidence but not with facility most social situations including introductions and casual conversations about current events, as well as work, family, autobiographical information; can handle limited work requirements, needing help in handling any complications or difficulties; can get the gist of most conversations on non-technical subjects (i.e. topics which require no specialized knowledge) and has a speaking vocabulary sufficient to express himself simply with some circumlocutions; accent, though often quite faulty, is intelligible; can usually handle elementary constructions quite accurately but does not have thorough or confident control of the grammar.
- 3 Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations on practical, social, and professional topics. Can discuss particular interests and special fields of competence with reasonable ease; comprehension is quite complete for a normal rate of speech; vocabulary is broad enough that he rarely has to grope for a word; accent may be obviously foreign; control of grammar good; errors never interfere with understanding and rarely disturb the native speaker.
- 4 Able to use the language fluently and accurately on all levels normally pertinent to professional needs. Can understand and participate in any conversation within the range of his experience with a high degree of fluency and precision of vocabulary; would rarely be taken for a native speaker, but can respond appropriately even in unfamiliar situations; errors of pronunciation and grammar quite rare; can handle informal interpreting from and into the language.

- 5 Speaking proficiency equivalent to that of an educated native speaker. Has complete fluency in the language such that his speech on all levels is fully accepted by educated native speakers in all of its features, including breadth of vocabulary and idiom, colloquialisms, and pertinent cultural references.

Factors in Speaking Proficiency

Pronun- ciation	Frequently not intelligible	Generally intelligible though constantly foreign	Generally intelli- gible, sometimes foreign	Always intelligible, occasionally slightly foreign	Indistinguish- able from native speaker
Grammar	Accuracy limited to set expressions; al- most no control of syntax; often con- veys wrong infor- mation	Fair control of most basic syntactic pat- terns; conveys mean- ing accurately in simple sentences most of time	Good control of most basic syntactic pat- terns; always con- veys meaning accu- rately in reasonably complex sentences	Makes only occasional errors, and these show no pattern of deficiency	Control equal to that of an educated na- speaker
Vocabulary	Adequate only for survival, travel, and basic courtesy needs	Adequate for simple social conversation and routine job needs	Adequate for parti- cipation in all general conversation and for professional discussions in a special field	Professional and general vocabulary broad and precise, appropriate to occasion	Equal to vo- cabulary of an educated native speaker
Fluency	Except for memo- rized expressions, every utterance required enormous obvious effort	Usually hesitant; often forced to silence by limita- tions of grammar and vocabulary	Rarely hesitant; always able to sus- tain conversation through circum- locutions	Speech on all pro- fessional matters as apparently effortless as in English; always easy to listen to	Speech at least as fluent as in English on all occasions
Comprehension	May require much repetition, slow rate of speech; understands only very simple, short familiar utterances	In general under- stands non-technical speech directed to him, but sometimes misinterprets or needs utterances re- worded. Usually cannot follow conversation between native speakers	Understands most of what is said to him; can follow speeches, clear radio broadcasts, and most conver- sation between native speakers, but not in great detail	Can understand all educated speech in any moderately clear context; oc- casionally baffled by colloquialisms and regionalisms	Equal to that of the native speaker
	73				74

Appendix E
FSI "Descriptive Statements"
Rating Criteria

PRONUNCIATION

- ☐ Indistinguishable from that of a native speaker.
- ☐ Occasional non-native pronunciation errors.
- ☐ Some phonemic errors, always intelligible.
- ☐ Frequent phonemic errors, generally intelligible.
- ☐ Frequent phonemic errors, often not intelligible.

GRAMMAR

- ☐ No grammatical errors made during the entire interview; or, if any errors are made, they are all of a type that could be made inadvertently by an educated native speaker.
- ☐ Sporadic grammatical errors of a type that would not be made by an educated native speaker. However, these are all of a rare or picayune nature.
- ☐ Occasional grammatical errors, generally good control in all areas.
- ☐ Frequent errors, but generally good control of basic constructions.
- ☐ Repeated errors in basic constructions.
- ☐ Virtually no grammatical control except in stock phrases.

VOCABULARY

- ☐ For any topic broached, vocabulary as extensive and precise as that of an educated native speaker who is not a specialist in the area being discussed.
- ☐ Vocabulary adequate to discuss any proposed topic, but occasional slight imprecision in choice of words or use of idioms.
- ☐ Vocabulary permits continuous conversation on a variety of topics, with some circumlocution.
- ☐ Limitations of vocabulary hinder discussion of some general topics.
- ☐ Vocabulary restricted to stock phrases and basic survival terms.
- ☐ Vocabulary inadequate for even the simplest conversation.

FLUENCY

- ☐ Speech is as smooth and as effortless as that of a native speaker.
- ☐ Speech shows only occasional evidence of "non-native" pausing.
- ☐ Numerous non-native pauses in course of interview.
- ☐ Speech is so halting and fragmentary that conversation is virtually impossible.

COMPREHENSION

- ☐ Comprehension equivalent to that of an educated native speaker.
- ☐ Understands everything in normal conversation except for very colloquial or low-frequency items or exceptionally rapid speech.
- ☐ Understands quite well educated speech spoken at a normal rate but requires occasional repetition or rephrasing.
- ☐ Deficiencies in comprehension require the interviewer to speak in a careful, somewhat simplified manner.
- ☐ Understands only slow, very simple speech on common social and personal topics; requires frequent repetition and rephrasing.
- ☐ Understands too little for even the simplest type of conversation.

Appendix F
Revised Speaking Test Scoring Key

AUTOBIOGRAPHICAL QUESTIONS

- 0 = no response or fails to communicate
- 1 = communicates, but with grammatical error(s)
- 2 = communicates, no grammatical error

READING ALOUD

Score entire passage for:

Pronunciation

- 0 = no response or only isolated words
- 1 = serious pronunciation problems make understanding difficult
- 2 = obviously non-native pronunciation but generally understandable
- 3 = only slight non-native characteristics, completely understandable
- 4 = completely native pronunciation

Fluency

- 0 = no response or only isolated words
- 1 = numerous (about 5 or more) unnatural pauses
- 2 = several (3-4) unnatural pauses
- 3 = few (1-2) unnatural pauses
- 4 = no unnatural pauses

SENTENCE REPETITION

- 0 = no response or only isolated words
- 1 = attempts to repeat sentence, but neither clause correct
- 2 = one clause repeated exactly
- 3 = both clauses repeated exactly

COMBINING SENTENCE ELEMENTS

- 0 = no response or no correct expansion
- 1 = some but not all expansions correct
- 2 = all expansions correct

GRAMMATICAL PATTERNS

- 0 = no response or response containing grammatical errors
- 1 = completely accurate response fulfilling requirements of pattern

SENTENCE COMPLETION

- 0 = no response or unmeaningful response
- 1 = meaningful response with grammatical error(s)
- 2 = meaningful response with no grammatical errors

DESCRIPTION OF PICTURE SEQUENCE

Separate score for:

Pronunciation

- 0 = no response or only isolated words
- 1 = serious pronunciation problems make understanding very difficult
- 2 = obviously non-native pronunciation but generally understandable
- 3 = only slight non-native characteristics, completely understandable
- 4 = completely native pronunciation

Grammatical Accuracy

- 0 = no response or only isolated words
- 1 = frequent grammatical errors impeding comprehension
- 2 = occasional grammatical errors
- 3 = very few errors, of a minor nature
- 4 = no grammatical errors

Fluency

- 0 = story not attempted or virtually no response
- 1 = more extensive than 0, but clearly hesitant
- 2 = somewhat hesitant
- 3 = reasonably smooth
- 4 = quite smooth

MULTIPLE QUESTIONS ON SINGLE PICTURE

For each question, separate score for:

Grammatical Accuracy

- 0 = no response or only isolated words
- 1 = grammatical errors
- 2 = no grammatical errors

Fluency

- 0 = no response or only isolated words
- 1 = hesitant response, obvious groping for words
- 2 = slight non-native disfluency
- 3 = native immediacy and ease of response

SIMULATED TELEPHONE CONVERSATION

For each response, separate score for:

Communication

- 0 = fails to convey message
- 1 = conveys message

Grammatical Accuracy

- 0 = no response or only isolated words
- 1 = grammatical errors
- 2 = no grammatical errors

Fluency

- 0 = no response or only isolated words
- 1 = hesitant response, obvious groping for words
- 2 = slight non-native disfluency
- 3 = native immediacy and ease of response

This Research Report is part of a series of reports on research relating to the Test of English as a Foreign Language. Other reports include:

The Performance of Native Speakers of English on the Test of English as a Foreign Language: Clark, John L.D. Report 1. November 1977.

Discusses the results of the administration of TOEFL to native speakers of English just prior to their graduation from a college-preparatory high school program. Total test score distributions were highly negatively skewed, reinforcing findings of earlier studies that TOEFL is not psychometrically appropriate for discriminating among native speakers of English with respect to English language competence.

An Evaluation of Alternative Item Formats for Testing English as a Foreign Language: Pike, Lewis W. Report 2. June 1979.

Describes an extensive research study conducted from 1972 to 1974 that was designed to explore possible changes in the format and content of TOEFL. Questions of validation, criterion selection, and content specifications were investigated. The report includes the results of these findings and discusses the implications for TOEFL content specifications and internal structure. This study contributed to the restructuring of TOEFL beginning in 1976.

The Performance of Non-Native Speakers of English on TOEFL and Verbal Aptitude Tests: Angelis, Paul J.; Swinton, Spencer S.; and Cowell, William R. Report 3. October 1979.

Gives the results of a study in which 400 graduate and undergraduate applicants took TOEFL, the GRE Verbal or the SAT Verbal, and the Test of Standard Written English. Included in the report are comparative data on performance across tests and interpretive information on how combined test results might best be used in the admission process.

The above reports are currently available. Other research reports are planned. For further information about any of the TOEFL Research Reports, write to:

TOEFL Program Office
Box 899
Princeton, NJ 08541, USA